Statistical Methods for High-Dimensional Networked Data Analysis

by

Yan Zhou

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2015

Doctoral Committee:

Professor Peter X. K. Song, Chair
Professor Matthias Kretzler
Assistant Professor Xiaoquan William Wen
Professor Ji Zhu

To my family

ACKNOWLEDGEMENTS

I would like to acknowledge many people for their help throughout my graduate study. I would first like to express my deepest gratitude to my advisor Dr. Peter X.-K. Song, for sharing his knowledge, encouraging me to work hard and constantly try to improve my work, and giving me the freedom to explore a diverse set of projects. He is the one who first sparked my interest in networked data analysis. I am not only respectful for his immense knowledge, motivation, and enthusiasm towards research, but also appreciate his invaluable guidance and enthusiastic encouragement during many difficult moments in my research. I could not have imagined having a better advisor and mentor for my Ph.D. study.

I would also like to thank my other thesis committee members Dr. Ji Zhu, Dr. Xiaoquan Willian Wen and Dr. Matthias Kretzler for their help and advices that made me more productive than what I could achieve on my own. All enlightening discussions I had with them have lead to better contents of this dissertation.

In addition, I would like to acknowledge my collaborators, Dr. Pei Wang, Dr. Betsy Lozoff and Dr. Fengji Geng, who offered me excellent opportunities and constructive suggestions to apply my statistical knowledge to solve real-world problems, which greatly strengthened the scientific background of my dissertation research.

I also want to take this opportunity to thank all the professors, staffs, my fellow graduate students and all my friends at the Department of Biostatistics, University of Michigan, for their support, encouragement and friendship.

Finally and most importantly, I would like to dedicate this thesis to my family

for their unconditional support and love, which set me worry free in my research endeavor.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Statistical Methods for High-dimensional Networked Data Analysis

by

Yan Zhou

Chair: Peter X. K. Song

Networked data are frequently encountered in many scientific disciplines. One major challenges in the analysis of such data are its high dimensionality and complex dependence. My dissertation consists of three projects.

The first project focuses on the development of sparse multivariate factor analysis regression model to construct the underlying sparse association map between gene expressions and biomarkers. This is motivated by the fact that some associations may be obscured by unknown confounding factors that are not collected in the data. I have shown that accounting for such unobserved confounding factors can increase both sensitivity and specificity for detecting important gene-biomarker associations and thus lead to more interpretable association maps.

The second project concerns the reconstruction of the underlying gene regulatory network using directed acyclic graphical models. My project aims to reduce false discoveries by identifying and removing edges resulted from shared confounding factors. I propose sparse structural factor equation models, in which structural equation models are used to capture directed graphs while factor analysis models are used to account for potential latent factors. I have shown that the proposed method enables

me to obtain a simpler and more interpretable topology of a gene regulatory network.

The third project is devoted to the development of a new regression analysis methodology to analyze electroencephalogram (EEG) neuroimaging data that are correlated among electrodes within an EEG-net. To address analytic challenges pertaining to the integration of network topology into the analysis, I propose hybrid quadratic inference functions that utilize both prior and data-driven correlations among network nodes into statistical estimation and inference. The proposed method is conceptually simple and computationally fast and more importantly has appealing large-sample properties. In a real EEG data analysis I applied the proposed method to detect significant association of iron deficiency on event-related potential measured in two subregions, which was not found using the classical spatial ANOVA random-effects models.

CHAPTER I

Introduction

1.1 Overarching goals

Massive complex data collected from biomedical studies presents a comprehensive set of significant analytic challenges in statistical modeling and data analysis. Many new opportunities emerge for the development of statistical methodologies useful to understand biological and disease mechanisms, leading to the development of better medical treatments and the improvement of patient's quality of life. The focus of my dissertation research is concerned with the development of new statistical models and data analytics to analyze high-throughput microarray data and large-scale neuroimaging data that arise from networks (e.g. genetic pathways and EEG electrodes on the scalp), and thus are correlated via networks.

More specifically my dissertation includes the following three projects:

Project I: Construction of high-dimensional sparse association maps between genes and biomarkers. In the first project of my dissertation, I focus on the development of an effective statistical approach to construct disease-related sparse genetic association maps with improved accuracy, through which I can not only segregate unobserved genetic variations from the noise but also detect master regulators, namely those genetic variants that are simultaneously correlated with multiple phenotypes (e.g. gene expressions).

Project II: Reconstruction of gene regulatory networks among genes. In the second project of my dissertation, I develop a new statistical procedure to construct sparse gene regulatory networks in the framework of Bayesian networks via directed acyclic graphical models. I utilize the factor analysis model to account for unobserved confounding, through which the proposed method can enable me to greatly improve both sensitivity and specificity in the discovery of causal relationships.

Project III: Regression analysis of networked data. In the third project of my dissertation, I develop a new regression analysis of networked data to assess potential adverse effects of prenatal exposure to iron deficiency on auditory recognition memory of two-month old infants, where memory functionality is measured by Electroencephalogram (EEG) neuroimaging. In this study, one of my primary interests is to incorporate certain established expert knowledge of brain functionality into statistical estimation and inference, so the resulting method enjoys better statistically powerful and meaningful discoveries.

The rest of this Chapter is organized as follows: Section 1.2 discusses some major challenges in high-throughput microarray data analysis and outlines the development of new methodology to construct genetic association maps for breast cancer. Section 1.3 highlights the importance of genetic regulatory network for the understanding of gene functions and cellular dynamics in system genomics. Some approaches of Bayesian network are reviewed, followed by an outline of my development of new methodology for a exploratory analysis of genetic pathway. Section 1.4 presents a brief introduction to the scientific background of prenatal iron deficiency on auditory recognition memory of newborn infants. Then, motivated by EEG neuroimaging data, this section discusses the formulation of regression analysis of networked data. Finally, Section 1.5 outlines the organization in the remainder of my dissertation.

1.2 Project I: Construction of association maps

1.2.1 Background

The currently available DNA microarray technology allows gene expression levels to be measured for the whole genome simultaneously across multiple samples. While the availability of genome-wide expression data has been increasing rapidly, the shortfall of relevant statistical techniques to analyze such high-throughout expression data is a clear concern. Alternative to uncovering single genes for complex traits, a system-based perspective is getting an increasing interest in elucidating the patterns underlying gene expression data and mechanisms related to the operation of a complex multicellular biology system. This gives rise to various opportunities for an enhanced understanding of functional genomics (*Allison et al.*, 2006; *Sieberts and Schadt*, 2007). However, a number of challenges arise from both the complexity of biological mechanism of genetic networks and a tremendously large number of genes/biomarkers, which create significant obstacles for understanding, analyzing and interpreting such massive high-dimensional data. In the recent literature, enormous efforts have been made to develop statistical methods for the analysis of high-throughput microarray data. In spite of much progress, there are still many gaps that demand new methodology development in order to furnish suitable statistical methods satisfying different needs in biomedical studies.

Important characteristics in the high-throughput microarray data analysis include the underlying gene-gene relationships and gene-trait relationships. In such analysis, a significant challenge often encountered in practice is that the sample size is small relative to the number of genes or biomarkers. A popular strategy to address this challenge is rooted in the utility of "sparsity" assumption, which refers to the scenario where the number of true signals is sparse and of lower dimension than the sample size. This assumption is widely used in many data mining approaches to unveil key

genetic features and network patterns of the underlying biological mechanism. Under the assumption of sparsity, statistical models will facilitate data analysis through effective data dimension reduction. Thus, the resulting model enables us to detect low-dimensional signals which is regarded as the most efficient strategy to analyze high-dimensional data.

There is a vast literature concerning the development of sparse models in the past two decades or so. I just name a few that are related closely to my project. *Tibshirani* (1996) first proposed the least absolute shrinkage and selection operator (LASSO) method based on L_1 norm penalty on regression coefficients, which has received a great deal of attention in the use of regularization techniques for variable selection and estimation. Since then, LASSO has been adapted or directly applied to a variety of statistical models and machine learning procedures, including but not limited to multivariate regression, generalized linear models, graphical models and principle component analysis (*Yuan and Lin*, 2007; *Zou*, 2006; *Zou et al.*, 2006; *Friedman et al.*, 2010a). In the meanwhile, the LASSO method has been also extended to accommodate different types of penalties on regression coefficients, such as group LASSO (*Yuan and Lin*, 2006), fused LASSO (*Tibshirani et al.*, 2005), elastic net (*Zou and Hastie*, 2005), nonnegative garrote (*Breiman*, 1995) and SCAD (*Fan and Li*, 2001). It is worth mentioning that some of these methods have been employed in high-throughput microarray data analysis, including, for example, gene regulatory network reconstruction (*Wille et al.*, 2004; *Shimamura et al.*, 2007), and genome-wide association studies (*Wu et al.*, 2009; *Kooperberg et al.*, 2010). More details may be found in a review paper by *Bansal et al.* (2010).

Genetic dependency structure among genes is another important characteristic of interest in the high-throughput microarray data analysis. In addition to the variations explained by some targeted genetic variants, there exists an extra dependence among gene expressions due to unobserved genetic or non-genetic factors (*Brem and*

Kruglyak, 2005). This is because, in general, a gene is likely to show strong expression correlations with other genes when they are in a common biological pathway and/or they share some measured and unmeasured genetic variants. It is noteworthy that dependencies among gene expressions may also be attributed to some shared non-genetic variants, such as environmental factors, population admixtures or kinship, batch effects in microarray experiments, changes in cellular composition, and other common physiological or biological factors. In order to examine and characterize such genetic dependency structure in high-throughput microarray data, I utilize the means of the factor analysis model, which has been extensively studied in the statistical literature, to capture potential latent factors attributive to dependencies among gene expressions.

The method of the factor analysis model is regarded as one of the most popular dimension reduction techniques, in which variations of correlated variables are modeled by a low number of latent factors. See for example, (*Kustra et al.*, 2006; *Stegle et al.*, 2008; *Friguet et al.*, 2009; *Blum et al.*, 2010), among others. Researchers have employed such model to deal with the genetic dependency structure in functional gene expression profiles. Also, *Carvalho et al.* (2008) pointed out that the pathway dependencies may be explained by some latent factors identified from the factor analysis model, which are further confirmed by some known biological structures. Thus, taking advantage of the factor model analysis, I am hopefully to develop an effective procedure to analyze pathway-specific gene expressions by dissecting them into co-regulated cellular mechanisms.

1.2.2 Motivating data

Project I is motivated by large datasets of RNA transcript levels and DNA copy numbers of about 20K genes/clones from more than 170 primary breast tumor specimens in a breast cancer cohort study. The primary aim is to conduct an integrative

analysis of DNA and RNA data that helps identify possibly more subtle (yet biologically important) genetic regulatory relationships in cancer cells. It is generally hard to construct either RNA-DNA associations or transcriptional regulatory networks with high accuracy due to a large proportion of masked signals. That is, in reality a sizable amount of measured gene expression variations are not only regulated by genetic variants of interest, but also possibly by other different genetic variants (e.g. microRNA regulations, DNA methylation) or even non-genetic variables (e.g. environmental exposures). Thus, it is critically important in the genetic association analysis to adjust confounding variables or the difference in cellular composition mentioned above, some of which may have been already measured but others are not measured in a study. A useful strategy to account for unmeasured confounding is to utilize the latent factor model to explicitly segregate the unmeasured confounding from the measurement noise. This idea has motivated Project I of my dissertation research.

1.2.3 Outline of methodology development

Gene expressions may be associated with copy number alterations (CNAs) in proximal genes within a several Mb window (cis-acting), as well as remote alterations throughout the genome (trans-acting). Simultaneously detecting genome-wide cis- and trans-acting associations is of primary interest in system biology because numerous passenger genes amidst the limited set of drivers may contribute to tumor progression. In addition, some CNAs, known as master regulators, play more important roles than other CNAs in the regulatory network, in terms of their ability of influencing many gene expressions simultaneously. Borrowing the sparse machine learning techniques, we may formulate the task of constructing a genetic regulatory association as a variable selection problem. Generalizing the classical LASSO regularization technique for the selection of individual predictors, I propose a double sparsity penalty function that encourages both group-wise and within-group sparsity

in Project I. As a result, the proposed method can select a master regulator that relates to a group of genes and can further detect non-zero associations of individual genes within an identified group including both cis-acting and trans-acting gene-CNA relationships. I hope to establish evidence that ignoring unmeasured confounding in such analysis may lead to not only the reduction of statistical power and true association signals, but also the loss of an opportunity to further explore latent genetic features related to the unmeasured confounding. Some of the existing methods such as principle component analysis (PCA) (*Wall et al.*, 2003) lack meaningful biological interpretations. The proposed statistical approach in Chapter 2, termed as *the sparse multivariate factor analysis regression model (smFARM)*, plays the following roles in the construction of sparse association maps: (i) identify master regulators; (ii) identify possible low-dimensional latent factors; and (iii) evaluate and interpret the impact of latent factors on the association map by a further gene-enrichment analysis.

1.3 Project II: Reconstruction of gene regulatory networks

1.3.1 Background

Reconstruction of gene regulatory networks (GRN) using gene expression data is of great importance in system biology as it pertains to the crucial knowledge of regulatory mechanisms in biological processes. A number of computational methods have been developed to infer gene networks from gene expression data. For example, co-expression or relevance network is constructed based on partial correlations in pairs of genes (*Butte et al.*, 2000; *Basso et al.*, 2005). Gaussian graphical model (GGM) has also been employed to construct gene networks (*Dobra et al.*, 2004; *Schfer and Strimmer*, 2005). However, the gaussian graphical model infers an undirected graph with edges being present (absent) if the corresponding pairs of genes are conditionally dependent (independent), given all other genes. In contrast, the gene regulatory

network is formulated as a type of causal network represented by a directed acyclic graph (DAG). DAG is also named as Bayesian network in the literature, which has been utilized to establish a dependency structure among genes (*Friedman et al.*, 2000; *Segal et al.*, 2003). To my best knowledge, *Xiong et al.* (2004) are the first to apply structural equation models (SEM) to gene regulatory network reconstruction using gene expression data. Recently, many machine learning methods have also been proposed to construct sparse DAGs, such as partial correlation (*Yang et al.*, 2011), regularized inverse covariance estimation (*Huang et al.*, 2006; *Levina et al.*, 2008), and sparse Bayesian network methods (*Li and Yang*, 2005; *Shojaie and Michailidis*, 2010; *Fu and Zhou*, 2013; *Aragam and Zhou*, 2014), among others. However, these existing methods often report many spurious findings, and an improvement is of great interest.

1.3.2 Motivating data

Project II is first motivated by my empirical insight obtained from Project I; that is, gene expression heterogeneity may be explained by both non-genetic latent factors and variables that characterize genetic pathways. The latter appears to be highly relevant to some biological functions. Therefore, without adjusting for latent factors, the analysis will lead to inflated false positive discoveries in the construction of gene regulatory network. Project II is motivated by a multivariate flow cytometry data given in *Sachs et al.* (2005), which has previously been analyzed by *Shojaie and Michailidis* (2010); *Fu and Zhou* (2013); *Friedman et al.* (2008), among others. This dataset includes 11 phosphorylated proteins and n=7466 cells. The consensus network, constructed by experimental annotations, has 20 directed edges and is used as the benchmark to assess the accuracy of an estimated network structure. According to *Shojaie and Michailidis* (2010), the ordering of the proteins in the consensus network is treated as prior information in the analysis.

1.3.3 Outline of methodology development

To reduce false discoveries, one strategy is to identify and remove those edges resulted from shared confounding factors. The key objective of Chapter 3 is to discern and quantify non-genetic variations from the gene expression measurements that are responsible for false causal relationships. Consequently, removing these non-genetic variations helps improve the construction of GRN. Given the annotated gene ordering (causality) in a pathway network, I develop *the structural factor equation model (SFEM)* by incorporating a factor analysis model into the structural equation model for a DAG, where the factor analysis model is to account for non-genetic confounders. The proposed SFEM may be converted into a generalized factor model, in which the model identifiability problem can be easily tackled. A LASSO-based penalized likelihood approach is developed for the SFEM to yield a sparse causal network. By comparing with the existing regularization methods such as the well-known PC algorithm, the proposed SFEM methodology outperforms in the case of small sample sizes, and therefore leads to simpler and more interpretable causal relationships in a GRN.

1.4 Project III: Regression analysis of networked data

1.4.1 Background

Iron deficiency (ID), one of the most common nutritional deficiencies in the world, is a public health challenge in both developing and developed countries. ID is the main cause of anemia, which affects 25% of the world's population (*McLean et al.*, 2009). Most of the affected populations are young children and women. In developing countries, the prevalence of ID is high in all age groups and demographics, especially among pregnant women and young children (*WHO*, 1999). Much attention has been gathered to reduce the prevalence of ID in young children, not only because of its

ubiquity, but also because many studies have revealed the negative effects of ID on individual development of social emotion, neuron behavior and cognition.

The fundamental scientific hypothesis that scientists are interested in is that these developmental challenges for ID infants and children might be due to ID-induced disrupted functioning in some brain areas, such as basal ganglia and hippocampus (*Lozoff and Georgieff*, 2006). For example, the hippocampus, important for learning, memory and the other cognitive functions, has been found to be vulnerable to ID during late prenatal and early postnatal period in animal models (*Fretham et al.*, 2011; *Lozoff and Georgieff*, 2006; *Radlowski and Johnson*, 2013; *Ranade et al.*, 2013). However, it is hard to directly assess the effect of prenatal and early postnatal ID on the development of hippocampus in human because of some ethical issues. The popular non-invasive approach is to use hippocampus-related cognitive functions, such as auditory or visual recognition memory, measured by electroencephalogram sensor net.

1.4.2 Motivating data

The electroencephalogram (EEG) data analyzed in Chapter 4 comes from one of our collaborative projects with scientists in the Center for Human Growth and Development, University of Michigan. Two-month old infant's brain electrical activity is collected during a period of 2000 milliseconds using a 64-channel EEG sensor net. The data collection occurs at two time points: when an infant hears his/her mother's voice and when hears a stranger's voice. At each time point, event-related potentials (ERP, a type of neuroimaging data), including P2, P750 and low slow wave (LSW), are recorded as primary outcomes of auditory recognition memory. These three ERPs are widely used as primary outcomes of auditory recognition memory (*Mai et al.*, 2012; *Siddappa et al.*, 2004). The scientific objective of the project is to evaluate whether or not, and if so how, iron deficiency affects auditory recognition memory for infants.

In this project, I consider the outcome LSW for illustration. Clearly, LSW measurements from 64 electrodes on an infant are correlated in the EEG-net, and such correlation is highly clustered according to subregions of memory functionality. According to our collaborators, dependence mechanisms of LSW measurements could be very complex that cannot be easily represented by conventional covariance or correlation matrices. For example, dependence symmetry may be invalid among electrodes, and strength of dependence may not be explicitly modeled due to the lack of legitimate distance metric bnetween electrodes. The standard analysis of the data using spatial ANOVA mixed-effects model (*Fields and Kuperberg*, 2012; *Gevins and Smith*, 2000) assumed implicitly symmetric exchangeable correlations among 64 nodes for the LSW data, and failed to detect significant association of iron deficiency on LSW. Thus, it motivates methodology research in Project III.

1.4.3 Outline of methodology development

In order to achieve high statistical efficiency in studying the effect of iron deficiency on infant's memory, I develop a method that strives to address a very important analytic challenge: to integrate some established knowledge of brain network topology into the estimation and inference for regression parameters. By recognizing the EEG-net as a network, I consider the marginal regression model for networked data in Chapter 4, because such model has great flexibility on allowing various forms of dependence structures among nodes and its ease on handling categorical outcomes. For the estimation of regression coefficients in the marginal model, both generalized estimating equation (GEE) (*Liang and Zeger*, 1986) and quadratic inference function (QIF) (*Qu et al.*, 2000) have been extensively studied in the literature. And many authors have advocated the importance of incorporating proper correlation structures in GEE or QIF to achieve desirable estimation efficiency; see for example, *Pan* (2001), *Qu et al.* (2008), *Wang and Carey* (2003) and *Zhou and Qu* (2012). However,

these two methods cannot be directly applied to deal with networked data because of the challenge on incorporating network dependence structures of potentially high dimension.

I develop a new method that allows combining two sources of knowledge regarding dependence structures of ERP amplitudes: one is the established or expert's prior knowledge about the subregions of functionality related to memory, and the other is the data-driven covariance from the available data at hand. In fact, I follow the strategy of *Stein* (1956)'s linear shrinkage estimation, which is later investigated by *Ledoit and Wolf* (2004) in the context of covariance matrix estimation. Since the shrinkage tuning parameter can be determined by maximizing the estimation efficiency, the proposed method is hoped to automatically allocate larger weights to more relevant correlation structures while to down weight non-informative structures. More importantly I establish large-sample properties in both estimation and inference for the proposed method. The proposed method is applied to detect significant association of iron deficiency on ERP measurements, which has not been found using the classical spatial ANOVA random-effects models.

1.5 Organization of the dissertation

The dissertation is structured as follows. Chapter 2 provides the detail concerning the development of *the sparse multivariate factor analysis regression model (smFARM)* to construct gene-biomarker association maps. In Chapter 3, I propose *the structural factor equation model (SFEM)* to reconstruct the gene regulatory network under the assumption that there is a natural ordering among nodes. Chapter 4 is devoted to the development of *regression analysis of networked data (RAND)* motivated by networked EEG neuroimaging data. I propose a new estimation method, termed as hybrid quadratic inference function (HQIF), for which related theoretical properties are established. Chapter 5 presents some concluding remarks and future

work. Most of technical details such as theoretical proofs are provided in appendices.

CHAPTER II

Sparse multivariate factor analysis regression models and its application to high-throughput array data analysis

In this chapter I present the sparse multivariate regression model that provides a useful tool to explore complex associations between multiple response variables and multiple predictors. When the multiple responses are correlated, ignoring such dependency will impair statistical power in the data analysis. Motivated by an integrative genomic data analysis, we propose a new methodology – sparse multivariate factor analysis regression model (smFARM), in which correlations of the response variables are analyzed by a factor analysis model with latent factors. This proposed method not only allows us to address the challenge that the number of regression parameters is larger than the sample size, but also to adjust for unobserved genetic and/or non-genetic factors that potentially conceals the underlying response-predictor associations. The proposed smFARM is implemented efficiently by utilizing the strength of the EM algorithm and the group-wise coordinate descend algorithm. The proposed methodology is evaluated and compared to the existing methods through extensive simulation studies. We apply smFARM in an integrative genomics analysis of a breast cancer dataset on the relationship between DNA copy numbers and gene expression

arrays to derive genetic regulatory patterns relevant to breast cancer.

2.1 Introduction

Unveiling regulatory patterns between genetic variants and gene expressions is of great importance to a broad range of biological studies, in the hope to improve our understanding of complex disease pathogenesis. As reported in many recent genetic studies, high-throughput gene expression array experiments and genotype or DNA copy number array experiments are carried out on the same set of subjects. This provides the unique opportunity to assess regulatory relationships among DNAs and RNAs. Copy number alterations (CNAs), including both germline variants and somatic copy number aberrations are found to be largely associated with disease mechanisms in many studies (*Pollack et al.*, 1999). In particular, somatic aberrations are discovered to be important for tumorigenesis. For example, oncogene activation by gene amplification or the loss of a tumor suppressor by gene deletion can cause transcriptional errors, which contributes to cancer pathogenesis (*Yuan et al.*, 2012). On the other hand, gene expression can be related to copy number alterations in proximal genes within a several Mb window (cis-acting), as well as remote alterations throughout the genome (trans-acting). It has been regarded as a difficult task to detect genomewide cis- and trans-acting effects simultaneously due to the fact that numerous passenger genes amidst the limited set of drivers may contribute to tumor progression. Recent studies (*Horlings et al.*, 2010; *Lahti et al.*, 2013; *Pollack et al.*, 2002) have focused on the cis-acting effects of copy number on gene expressions and there are few studies that have considered trans-acting effects on a genomewide scale. To address these challenges require new analytic tools suitable for well-powered genomic studies.

The construction of genome-wide regulatory map by exploiting genomic and transcriptomic data typically involves in a large number of gene expressions as response

variables and high-dimensional genetic variants (e.g. DNA copy number alterations) as predictors. This analytic task can be primarily formulated by a multivariate regression analysis (*Bedrick and Tsai*, 1994; *Lutz and Buhlmann*, 2006). Usually, the genetic regulatory relationships are intrinsically sparse, in the sense that one genetic variant may regulate only a small proportion of gene expressions, rather than the majority of them. It is also reported that some genetic variants, known as master regulators, play more important roles than other variants in the regulatory network, in terms of their ability of influencing many gene expressions simultaneously (*Gardner et al.*, 2003; *Jeong et al.*, 2001). Thus, it is of great interest to develop proper multivariate regression models that account for both the sparsity in the regulatory relationships and the existence of master regulators in the mapping of genetic associations. Towards this goal, sparse penalty functions such as LASSO (*Tibshirani*, 1996), elastic net (*Zou and Hastie*, 2005), and group LASSO (*Yuan and Lin*, 2006) have been introduced to the multivariate regression framework (e.g. *Lutz and Buhlmann* (2006); *Turlach et al.* (2005); *Yuan et al.* (2012)). Readers can find more details about the comparison of our work with the existing method in Section 2.5.

Some researchers (e.g. *Gibson* (2008); *Leek and Storey* (2007)) have pointed out that gene expressions are influenced by many biological and non-biological factors. Biological factors could include, for example, genotype polymorphisms/mutations, DNA copy number variations, DNA methylation, microRNA regulations, protein regulations and others. Non-biological factors include sample collection noises, instrumental errors, and batch effects. In addition, population admixtures or kinship in a study population may also influence data generation mechanism of gene expression profiles. Because of these complications, quite often only a small portion of variations in gene expressions can be explained by one type of genetic predictors under investigation. Moreover, it is reported that gene expression heterogeneity is presented strongly in many studies but it is not yet properly taken into account in

statistical analysis. For example, *Leek and Storey* (2007) and *Stegle et al.* (2008) have showed that gene expression heterogeneity not only leads to the reduction of statistical power but also produces spurious association signals when studying the regulatory relationships between genotypes and gene expressions. This motivates us to develop a new method that employs the factor analysis model to account for such heterogeneity attributed to some unobserved genetic and/or non-genetic variabilities. As a result, we can improve both statistical power and accuracy of identifying significant associations between genes and genetic markers.

In this chapter, we plan to achieve three objectives via a sparse multivariate factor analysis regression model (smFARM): (i) to identify both trans-acting and cis-acting effects in one modeling framework; (ii) to regularize the association map by encouraging the selection of important predictors; and (iii) to estimate the covariance matrix of the response variables via the means of multivariate factor analysis. The factor analysis model enables us to understand and interpret additional association features beyond what expression-genetic variant associations describe. The mean model component of smFARM is parameterized by a matrix of regression coefficients that are supposed to contain many zeros because of sparse genetic regulatory relationships. This part of modeling relates closely to the remMap method proposed by *Peng et al.* (2010) for the identification of genetic regulatory relationships and master predictors using a regularized multivariate regression model. Compared to remMap, our proposed smFARM further extends their model and captures residual correlations of the responses using latent factors. As discussed earlier, when studying the regulatory relationships between gene expressions and DNA copy numbers, gene expression levels could be often confounded by unobserved genetic and/or non-genetic factors. Thus, incorporating the latent factors in smFARM leads to a more efficient method to extract important features of the regulatory network than remMap. This is shown in our analysis of the same breast cancer data set, which was previously analyzed in

Peng et al. (2010). We find that smFARM is able to identify several new novel regulatory relationships between gene expressions and copy number alternation intervals (CNAIs). Another new contribution of important is the analysis and interpretation of the latent factors. By utilizing gene set enrichment analysis (GSEA), we decompose these latent factors into pathway subcomponents and reveal additional variations highly relevant to some important biological functions of breast cancer.

The remainder of this chapter is organized as follows. In Section 2.2, we first introduce multivariate factor analysis regression model (mFARM) and then develop the sparse regularization procedure in Section 2.3. Section 2.4 presents an efficient EM-GCD algorithm and its implementation. Also, criteria for selecting tuning parameters and the number of latent factors are discussed in Section 2.4. Section 2.5 discusses the relationship between our method and other available methods. Section 2.6 is devoted to the evaluation of the proposed approach through extensive simulation studies. Section 2.7 presents the analysis of breast cancer data to illustrate the application of the proposed smFARM. We provide some concluding remarks in Section 2.8. Related technical details are included in the Appendix A, B and C.

2.2 Model

2.2.1 Multivariate regression model

Multivariate regression model plays an important role in multivariate data analysis. Such model extends the classical one-dimensional regression model, which is widely used to deal with correlated response variables. Following the common notations in multivariate regression model, for subject i, we assume that the conditional distribution of a $Q\times 1$ random vector $y_i = (y_{i1}, \ldots, y_{iQ})^T$ given P-element explanatory vector $x_i = (x_{i1}, \ldots, x_{iP})^T$ is a multivariate normal distribution. And its expectation

is specified by the following linear equations:

$$E(y_i|x_i) = \Theta x_i,\ i = 1, \ldots, N, \tag{2.1}$$

where $\Theta = \{\theta_{qp}\}$ is a $Q \times P$ matrix of unknown regression coefficients, and its covariance is $\mathrm{Var}(y_i|x_i) = \Sigma$, which is an unknown $Q \times Q$ positive definite covariance matrix independent of x_i. Obviously, if $Q = 1$, model (2.1) becomes the classical one-dimensional regression model, where Θ is a P-dimensional regression coefficient vector. In matrix Θ, the q-th row represents the vector of regression coefficients corresponding to the q-th regression model, i.e. $E(y_{iq}|x_i) = \sum_{p=1}^{P} \theta_{qp} x_{ip}$, which regresses the q-th response variable y_{iq} on all P predictors. Clearly, the ordinary least square method (or equivalently the maximum likelihood method under the normally distributed errors) yields the estimator of Θ as $\hat{\Theta}^T = (X^T X)^{-1} X^T Y$. This implies that each row of Θ can be estimated separately by regressing each of Q responses on the P predictors ignoring the dependence across the Q responses. This is because in this estimation there are no common coefficients and/or common parameters in Σ shared across Q individual one-dimensional regression models. In contrast, when some common features are present in the mean models and/or covariance matrices, borrowing strengths across different margins will be beneficial. Consequently, joint estimation involving all Q rows would provide better statistical power.

2.2.2 Factor analysis model

In this section, we propose to model the covariance Σ by the following factor analysis model:

$$\Sigma = BB^T + \Psi, \tag{2.2}$$

where B is a $Q \times K$ matrix of factor loadings pertinent to communalities for K ($\leq Q$) latent factors and Ψ is a $Q \times Q$ diagonal matrix of uniqueness. Clearly, the mean

model (2.1) does not involve the K latent factors, while the covariance model (2.2) is determined by loadings B and uniqueness Ψ. Factor analysis is one of the popular dimension reduction techniques that represents variations of correlated variables by a low number of latent factors. See for example, *Blum et al.* (2010), *Friguet et al.* (2009), *Kustra et al.* (2006) and *Stegle et al.* (2008), among others, in which the factor analysis model has been employed to deal with heterogeneity in functional gene expression profiles.

2.2.3 Multivariate factor analysis regression model

Combining models (2.1) and (2.2), with P predictors x_i and K unobserved latent factors $z_i = (z_{i1}, \ldots, z_{iK})^T$, we propose the following multivariate factor analysis regression model (mFARM):

$$y_i = \Theta x_i + B z_i + \epsilon_i, \ i = 1, \ldots, N, \tag{2.3}$$

where z_i's are i.i.d. K-variate vectors of latent factors following multivariate normal distribution $\text{MVN}_K(0, I)$, and ϵ_i's are i.i.d. measurement errors with $\text{MVN}_Q(0, \Psi)$ and are independent of the latent factors $z_{i1}, \ldots, z_{iK}$. In matrix notation, model (2.3) may be rewritten as follows:

$$Y = X\Theta^T + ZB^T + E, \tag{2.4}$$

where $Y^T_{Q\times N} = (y_1, \ldots, y_N)$, $X^T_{P\times N} = (x_1, \ldots, x_N)$, $Z^T_{K\times N} = (z_1, \ldots, z_N)$ and $E^T_{Q\times N} = (\epsilon_1, \ldots, \epsilon_N)$. For simplicity, we assume that all Q responses and all P predictors are standardized to have zero mean and thus the intercept terms are removed from (2.4).

Our proposed mFARM model (2.4) will improve the capacity of statistical analysis for the construction of genetic regulatory maps with high-throughput array data, because it accounts for unobserved factors that better capture variabilities in the

residuals.

2.3 Regularized Estimation

To achieve sparsity in the estimation of parameter matrix Θ, which characterizes the association map of interest, and to encourage the detection of master predictors (i.e. master regulators) in a similar spirit to the remMap method (*Peng et al.*, 2010), we propose the following doubly penalized loss function:

$$\begin{aligned} L(\Theta, \Psi, B) =& \frac{1}{2N} \sum_{i=1}^{N} (y_i - \Theta x_i)^T (BB^T + \Psi)^{-1} (y_i - \Theta x_i) \\ &+ \lambda_1 \sum_{q=1}^{Q} \sum_{p=1}^{P} |\theta_{qp}| + \lambda_2 \sum_{p=1}^{P} \sqrt{\theta_{1p}^2 + \cdots + \theta_{Qp}^2}, \end{aligned} \tag{2.5}$$

where λ_1 and λ_2 are two nonnegative tuning parameters. The first L_1 norm penalty term in the above loss function controls the overall sparsity in Θ, while the second L_2 norm penalty term controls the column sparsity in Θ. The use of two penalties facilitates the selection of important predictors that affect multiple responses simultaneously.

If there is some *a priori* knowledge about the known relationship between a predictor X_p and a response Y_q, such information may be incorporated into the optimization procedure in a similar way suggested in *Peng et al.* (2010). That is, consider a prespecified $Q \times P$ matrix C^* whose (q, p)-th element as:

$$C_{qp}^* = \begin{cases} 2, \text{ if } X_p \text{ is independent of } Y_q; \\ 0, \text{ if } X_p \text{ is associated with} Y_q; \\ 1, \text{ if there is no prior information.} \end{cases} \tag{2.6}$$

As a result, given an unknown matrix Θ^*, the (q, p)-th entry θ_{qp}^* will be set as 0 in advance if $C_{qp}^* = 2$; otherwise, θ_{qp}^* will or will not be penalized according to $C_{qp}^* = 1$

or $C^*_{qp} = 0$. After setting matrix $\Theta = \Theta^*$ according to C^*, the modified objective function is given by

$$\begin{aligned} L(\Theta, \Psi, B) = & \frac{1}{2N} \sum_{i=1}^{N} (y_i - \Theta x_i)^T (BB^T + \Psi)^{-1} (y_i - \Theta x_i) \\ & + \lambda_1 \sum_{q=1}^{Q} \sum_{p=1}^{P} |C_{qp} \theta_{qp}| + \lambda_2 \sum_{p=1}^{P} \sqrt{C_{1_p} \theta_{1p}^2 + \cdots + C_{Qp} \theta_{Qp}^2}, \end{aligned} \tag{2.7}$$

where a $Q \times P$ matrix $C = \{C_{qp}\}$ is defined as $C_{qp} = 1\{C^*_{qp} = 1\}$.

Without loss of generality, we assume that both λ_1 and λ_2 are positive, and if one of them is zero, we can modify our methodology with little effort. Also, the proposed smFARM may be used to deal with the case of high-dimensional measurements with $min(P, Q) \gg N$, which is pervasive in biological studies, such as microarray data that contain thousands of biological markers from typically dozens to hundreds of subjects.

2.4 EM-GCD Algorithm

In this section, we estimate three unknown parameter matrices, (Θ, B, Ψ), through minimizing the doubly penalized loss function (2.7), where Θ and (B, Ψ) are involved in the mean model and the covariance model, respectively. In this section, we propose a two-step iterative approach to estimate these three matrices. Given the current estimates of $(B^{(t)}, \Psi^{(t)})$, $\Theta^{(t+1)}$ is given through minimizing the doubly penalized loss function (2.7), and $(B^{(t+1)}, \Psi^{(t+1)})$ are updated through the EM algorithm presented below in Section 2.4.1. Repeating these two-step procedure iteratively till convergence, we obtain estimates $(\hat{\Theta}, \hat{B}, \hat{\Psi})$ in the end.

Before introducing our formulation, we summarize some notations used in this section. Let I_K, I_Q and I_N denote identity matrices with dimension K, Q and N, respectively. Given (q_0, p_0) as an arbitrary target for updating, we denote A_{p_0} as the

p_0-th column of a matrix $A \in \mathbb{R}^{Q\times P}$ and $A_{q_0p_0}$ as the q_0-th element of A_{p_0}. $\|A_{p_0}\|_0$, which counts a total number of non-zero elements in a vector, is the L_0 norm of A_{p_0}. Notation $[A_{p_0}(q_0)]$ is a vector where $[A_{p_0}(q_0)]_q = A_{qp_0}$, if $q \neq q_0$; and $[A_{p_0}(q_0)]_{q_0} = 0$, otherwise. And notation $[A(\cdot, p_0)]$ is a matrix where $[A(\cdot, p_0)]_p = A_p$, if $p \neq p_0$; and $[A(\cdot, p_0)]_{p_0} = 0$, otherwise. Similarly, $[A(q_0, p_0)]$ is a matrix where $[A(q_0, p_0)]_{qp} = A_{qp}$, if $(q, p) \neq (q_0, p_0)$; and $[A(q_0, p_0)]_{q_0p_0} = 0$, otherwise.

2.4.1 EM algorithm

The EM algorithm is used to estimate (B, Ψ) in the factor analysis model. Note that, there are several well established methods to estimate these factor model parameters, such as principal component estimation method, maximum likelihood estimation method (*Johnson and Wichern*, 2007) and EM algorithm (*Rubin and Thayer*, 1982). Under the normality assumption, we may implement the EM algorithm by treating the latent factors z_i as "missing data" and Θ as a fixed "known" constant matrix and then maximizing the joint log-likelihood of the full data $\{(y_i^* \triangleq y_i - \Theta x_i,\ z_i),\ i = 1, \ldots, N\}$. In fact, using the EM algorithm to estimate Ψ, B and Z enjoys computational simplicity and numerical stability. This is because the algorithm only operates the matrix inverse, $(BB^T + \Psi)^{-1} = \Psi^{-1} - \Psi^{-1}B(B^T\Psi^{-1}B + I_K)^{-1}B^T\Psi^{-1}$, at low-dimension K, instead of high-dimension Q, which gives rise to computational efficiency in the parameter estimation. It is easy to derive the EM algorithm updates of B and Ψ at the $(t+1)$-th iteration, respectively, given by

$$B^{(t+1)} = \Big\{ \sum_{i=1}^{N} y_i^* E\Big(z_i^T | y_i^*; B^{(t)}, \Psi^{(t)}\Big) \Big\} \Big\{ \sum_{i=1}^{N} E\Big(z_i z_i^T | y_i^*; B^{(t)}, \Psi^{(t)}\Big) \Big\}^{-1},$$

$$\Psi^{(t+1)} = \frac{1}{N} diag\Big\{ \sum_{i=1}^{N} y_i^* y_i^{*T} - B^{(t+1)} \sum_{i=1}^{N} E\Big(z_i | y_i^*; B^{(t)}, \Psi^{(t)}\Big) y_i^{*T} \Big\}.$$

Specifically, if $\Psi = \sigma^2 I_Q$ in the case of LOw-Rank representation and Sparse regression (LORS) (*Yang et al.*, 2013), we consider a simple re-parameterization by letting $\tilde{B} = \sigma^{-1}B$, and $\tilde{z}_i = \sigma z_i$. Clearly, Bz_i and $\tilde{B}\tilde{z}_i$ follow the same distribution. Thus, the EM algorithm updates $\tilde{B}$ and σ^2 at the $(t+1)$-th iteration by the following expressions:

$$\begin{aligned}
\tilde{B}^{(t+1)} &= \Big\{ \sum_{i=1}^{N} y_i^* E\Big(\tilde{z}_i^T | y_i^*; \tilde{B}^{(t)}, \sigma^{2(t)}\Big)\Big\}\Big\{ \sum_{i=1}^{N} E\Big(\tilde{z}_i \tilde{z}_i^T | y_i^*; \tilde{B}^{(t)}, \sigma^{2(t+1)}\Big)\Big\}^{-1}, \\
\sigma^{2(t+1)} &= \frac{1}{NQ} \sum_{i=1}^{N} y_i^{*T} \Big\{ I_Q - \tilde{B}^{(t)}\big(I_K + \tilde{B}^{T(t)}\tilde{B}^{(t)}\big)^{-1}\tilde{B}^{T(t)}\Big\} y_i^*.
\end{aligned}$$

2.4.2 Group-wise coordinate descent (GCD) algorithm

We implement an efficient algorithm to yield the optimal solution to minimizing (2.7) under a fixed positive definite $\Sigma \stackrel{\text{def}}{=} BB^T + \Psi$, along the lines of the sparse group LASSO (*Friedman et al.*, 2010b). Since minimizing (2.7) with respect to Θ is equivalent to a convex optimization problem, the objective function decreases over iterations, and the algorithmic convergence is warranted (*Tseng*, 2009). A similar algorithm for the loss function (2.7) with the adaptive L_1 norm and L_2 norm penalties can be established with minor modifications on the one presented below.

Given C_{p_0}, we first define an "inactive" set $\mathscr{A}_{p_0} \stackrel{\text{def}}{=} \{q : C_{qp_0} = 0, 1 \le q \le Q\}$, and an "active" set $\mathscr{B}_{p_0} \stackrel{\text{def}}{=} \{q : C_{qp_0} = 1, 1 \le q \le Q\}$. $\theta_{p_0} = (\theta_{1p_0}, \ldots, \theta_{Qp_0})^T$ is split into two sub-vectors $\theta_{p_0}^{\mathscr{A}}$ and $\theta_{p_0}^{\mathscr{B}}$, where $\theta_{qp_0}^{\mathscr{A}} = \theta_{qp_0}$, if $q \in \mathscr{A}_{p_0}$; and $\theta_{qp_0}^{\mathscr{A}} = 0$, otherwise. Similarly, $\theta_{qp_0}^{\mathscr{B}} = \theta_{qp_0}$, if $q \in \mathscr{B}_{p_0}$; and $\theta_{qp_0}^{\mathscr{B}} = 0$, otherwise. Given $R_{p_0} = \|C_{p_0}\|_0$, $H_{\mathscr{A}_{p_0}}$ is an $R_{p_0} \times Q$ matrix of full rank with elements 0 or 1, satisfying $H_{\mathscr{A}_{p_0}}\theta_{p_0}^{\mathscr{A}} = 0$, which sets $\theta_{qp_0}^{\mathscr{A}} = 0$ for $q \notin \mathscr{A}_{p_0}$. Finally, let $S(\cdot,\ \cdot)$ denote the soft-thresholding operator $S(z,\ \gamma) = sgn(z)(|z| - \gamma)_+$. The following proposition gives a procedure for sequentially updating Θ one column at a time.

Proposition II.1. *Given* $[\Theta(\cdot, p_0)]$, $\hat{\theta}_{p_0} = \hat{\theta}_{p_0}^{\mathscr{A}} + \hat{\theta}_{p_0}^{\mathscr{B}}$, *a minimizer of (2.7), satisfies:*

(i) Suppose $\theta^{\mathscr{B}}_{p_0}$ *is withheld. Let* $\widetilde{Y}^T_{\mathscr{A}_{p_0}} = Y^T - [\Theta(\cdot, p_0)]X^T - \theta^{\mathscr{B}}_{p_0}X^T_{p_0}$*, we have*

$$\hat{\theta}^{\mathscr{A}}_{p_0} = \Big\{I_Q - \Sigma H^T_{\mathscr{A}_{p_0}}\Big(H_{\mathscr{A}_{p_0}}\Sigma H^T_{\mathscr{A}_{p_0}}\Big)^{-1}H_{\mathscr{A}_{p_0}}\Big\}\widetilde{Y}^T_{\mathscr{A}_{p_0}}X_{p_0}/\big\|X_{p_0}\big\|_2^2, \tag{2.8}$$

(ii) Suppose $\theta^{\mathscr{A}}_{p_0}$ *is withheld. Let* $\widetilde{Y}^T_{\mathscr{B}_{p_0}} = Y^T - [\Theta(\cdot, p_0)]X^T - \theta^{\mathscr{A}}_{p_0}X^T_{p_0}$*, we have*

$$\hat{\theta}^{\mathscr{B}}_{p_0} = 0, \ \textit{if} \ \Big\|\frac{1}{N}\Sigma^{-1}\widetilde{Y}^T_{\mathscr{B}_{p_0}}X_{p_0} - \lambda_1 s_{p_0}\Big\|_2 \le \lambda_2, \tag{2.9}$$

where $s_{p_0} = (s_{1p_0}, \ldots, s_{Qp_0})^T$*, with*

$$s_{qp_0} \ = \ \begin{cases} sgn(\hat{\theta}_{qp_0}), & \textit{if} \ \ \hat{\theta}_{qp_0} \neq 0 \ \textit{and} \ q \in \mathscr{B}_{p_0}, \\ \in [-1, 1], & \textit{if} \ \ \hat{\theta}_{qp_0} = 0 \ \textit{and} \ q \in \mathscr{B}_{p_0}, \\ 0, & \textit{if} \ \ q \notin \mathscr{B}_{p_0}; \end{cases}$$

(iii) If $\hat{\theta}^{\mathscr{B}}_{p_0}$ *does not satisfy condition (2.9), given* $[\Theta(q_0, p_0)]$ *and* $\theta^{\mathscr{B}}_{p_0}$ *obtained from current estimates of* Θ*,* $\theta_{q_0p_0}$ *with* $q_0 \in \mathscr{B}_{p_0}$ *is estimated by*

$$\hat{\theta}_{q_0p_0} = \begin{cases} \dfrac{NS\Big(\frac{1}{N}X^T_{p_0}\big(Y - X[\Theta(q_0, p_0)]^T\big)\Sigma^{-1}_{q_0}, \ \lambda_1 + \lambda_2\Big)}{\Sigma^{-1}_{q_0q_0}\big\|X_{p_0}\big\|_2^2}, \ \textit{if} \ \big\|[\theta^{\mathscr{B}}_{p_0}(q_0)]\big\|_2 = 0, \\ 0, \ \textit{if} \ \Big|\dfrac{1}{N}X^T_{p_0}\big(Y - X[\Theta(q_0, p_0)]^T\big)\Sigma^{-1}_{q_0}\Big| \le \lambda_1, \ \textit{and} \ \big\|[\theta^{\mathscr{B}}_{p_0}(q_0)]\big\|_2 \neq 0, \\ \dfrac{NS\Big(\frac{1}{N}X^T_{p_0}\big(Y - X[\Theta(q_0, p_0)]^T\big)\Sigma^{-1}_{q_0}, \ \lambda_1\Big)}{\Sigma^{-1}_{q_0q_0}\big\|X_{p_0}\big\|_2^2 + 2N\lambda_2\big\|\theta^{\mathscr{B}}_{p_0}\big\|_2^{-1}}, \ \textit{otherwise.} \end{cases} \tag{2.10}$$

From Proposition II.1, we know that if Σ is an identity matrix, $\hat{\theta}_{q_0p_0}$ with $q_0 \in \mathscr{A}_{p_0}$ given in (2.8) is actually an ordinary least square estimate. In addition, (2.10) gives us the coordinate descent method to update Θ. To sum up, our procedure in Proposition II.1 can efficiently shrink a group of "active" predictors to exactly zero by (2.9), while also shrink some individuals within that group to zero. The detailed proof of Proposition II.1 is given in the Appendix B. The GCD algorithm is implemented by

the following steps:

Algorithm 1 GCD algorithm

Step 1. Start with an initial value $\hat{\Theta} = \Theta^{(0)}$.
Step 2a. Given p_0-th column of Θ, $\hat{\theta}^{\mathscr{A}}_{p_0}$ is updated via (2.8).
Step 2b. Check if (2.9) is satisfied. If so, set $\hat{\theta}^{\mathscr{B}}_{p_0} = 0$.
Step 2c. If (2.9) is not satisfied, update $\hat{\theta}_{q_0 p_0}$ for $q_0 \in \mathscr{B}_{p_0}$ via (2.10).
Step 2d. Step 2c is iterated until convergence.
Step 3. Iterate the entire subloop of Step 2 over $p_0 = 1, \ldots, P$ until convergence.

Finally, a combination of the EM algorithm and GCD algorithm, termed as the EM-GCD algorithm in this section, allows us to iteratively update Θ, B and Ψ. The detail is provided in the following Algorithm 2.

Algorithm 2 EM-GCD algorithm

Step 1. Set an initial value $\Theta^{(0)}$ whose $(q,\ p)$-th element is

$$\hat{\theta}^{(0)}_{qp} = \begin{cases} \dfrac{NS(\frac{1}{N}X_p^T Y_q,\ \lambda_1)}{\|X_p\|_2^2}, \text{ if } C_{qp} = 1, \\ X_p^T Y_q / \|X_p\|_2^2, \text{ if } C_{qp} = 0. \end{cases}$$

Let $\Psi^{(0)} = I_Q$, and let $B^{(0)}$ be the first K right-singular vectors of $Y - X\Theta^{(0)^T}$.
Step 2a. Given $B^{(t)}$, $\Psi^{(t)}$, and $\Theta^{(t)}$, at iteration t+1, $\Theta^{(t+1)}$ is updated by the GCD algorithm that sequentially updates one column θ_p $(p = 1, \ldots, P)$ of Θ at a time, until convergence.
Step 2b. Given $\Theta^{(t+1)}$, update $B^{(t+1)}$ and $\Psi^{(t+1)}$ iteratively using the EM algorithm till convergence.
Step 3. Repeat the two-step cycle, 2a and 2b, until convergence.
Step 4. Output the final estimates $\hat{\Theta}$, $\hat{B}$ and $\hat{\Psi}$.

2.4.3 Tuning parameter selection

We first consider the selection of the tuning parameters (λ_1, λ_2) with a given $K = K_0$, and then discuss the selection of K. Following *Peng et al.* (2010), we adopt the M-fold cross-validation method to choose the tuning parameters $(\lambda_1,\ \lambda_2)$. Since the true model is believed to be sparse as suggested by *Peng et al.* (2010), we utilize the ordinary least squares (OLS) estimates instead of the shrunken estimates

to calculate the cross-validation score. This is because, when there are many potential poor predictors, the cross-validation score based on shrunken estimates often leads to severe false positive rates (*Peng et al.*, 2010; *Efron et al.*, 2004). In contrast, using the OLS estimates seems to make a reasonable remedy for such a problem, which is also observed in our simulation studies. It is worth pointing out that Bayesian information criterion (BIC), another popular tuning selection method, is not considered here, mainly because estimating the degrees of freedom needed in the BIC is difficult under a nonorthogonal design.

We now turn to discuss the selection of the number of latent factors K. The number of latent factors in the proposed smFARM can affect the resulting sparsity of regression coefficient matrix Θ and have to be tuned properly. Basically, selecting the number K can be implemented in the M-fold cross validation procedure. For more details, see Appendix A. In this section, for the computational ease, BIC is also applied to choose the number K. The value of BIC with K latent factors is given by:

$$\text{BIC}(K) = \log\left\{N^{-1}\|Y - X\widetilde{\Theta}^T - \widehat{E(Z|Y)}\hat{B}^T\|_F\right\} + \log(N)\widehat{df}(K)/N, \qquad (2.11)$$

where $\widetilde{\Theta}$ and $(\hat{B}, \widehat{E(Z|Y)})$ are obtained from a re-estimation step based on OLS model and the EM algorithm discussed in Sections 2.4.1 and 2.4.2. Here $\|\cdot\|_F$ denotes the Frobenius norm of matrix.

To determine degrees of freedom in (2.11), we note that the factor analysis model can be viewed as a special nonlinear smoothing procedure of the form: $\hat{Y} = (I_N - \widehat{H_Z})X\widetilde{\Theta}^T + \widehat{H_Z}Y$, where $\widehat{H_Z} = E(\widehat{Z|Y})\{E(\widehat{ZZ^T|Y})\}^{-1}E(\widehat{Z|Y})^T$ stands for the smoothing matrix. Thus, the degrees of freedom $\widehat{df}(K) = tr[\partial\hat{Y}/\partial Y|\widetilde{\Theta}]$ may be approximated by $tr[\widehat{H_Z}]$.

In summary, Algorithm 3 presents the steps to select tuning parameters λ_1, λ_2, and K.

Algorithm 3 Tuning parameter selection

Step 1a. Given the number of latent factors $K = K_0$, select the optimal λ_1^* and λ_2^* by the M-fold cross validation method.

Step 1b. Re-estimate $\widetilde{\Theta}(\lambda_1^*, \lambda_2^*)$ from the OLS regression as well as $\hat{B}$ and $\widehat{E(Z|Y)}$ from the EM algorithm.

Step 1c. Calculate BIC at $K = K_0$ using (2.11).

Step 2. Letting K vary from 0 to a given large number, select the optimal K that minimizes the BIC.

2.5 Relationship to the existing methods

In the past two decades or so, many regularized variable selection methods have been proposed in the statistical literature, including but not limited to LASSO (*Tibshirani*, 1996), group LASSO (*Yuan and Lin*, 2006), fused LASSO (*Tibshirani et al.*, 2005), elastic net (*Zou and Hastie*, 2005), nonnegative garrote (*Breiman*, 1995), and SCAD (*Fan and Li*, 2001). Some of these methods or their variants have been specifically developed in the context of multivariate regression models. For example, *Turlach et al.* (2005) considered the max-L_1 penalty to select a common subset of predictors in multiple response regression; *Yuan and Lin* (2007) proposed a dimension reduction method by encouraging sparsity among singular values in the regression coefficient matrix; *Peng et al.* (2010) developed a regularized method to identify master predictors via a mixture of L_1 and L_2 penalties. However, all the existing methods have not addressed structures of dependencies among multiple responses, which potentially leads to the loss of an opportunity to further explore potentially important features contained in the residuals.

Recently, *Rothman et al.* (2010) proposed multivariate regression with covariance estimation (MRCE) method, a penalized log-likelihood approach with L_1 penalty to select a subset of predictors while accounting for correlated errors. Later both *Lee and Liu* (2012) and *Yin and Li* (2011) extended MRCE (*Rothman et al.*, 2010) to explore the conditional independence relationships among responses via covariance matrix Σ, adjusting for possible genetic effects on gene expressions. Following the

same idea, *Cai et al.* (2013) presented a covariate-adjusted precision matrix estimation (CAPME) method. Different from the approaches proposed by *Lee and Liu* (2012), *Rothman et al.* (2010) and *Yin and Li* (2011), the CAPME approach does not make the multivariate normal assumption on the error distribution. It is noted that all the existing methods have concerned only with constructing a conditional association network among the responses (e.g. gene-gene relationship), instead of focusing on constructing a response-predictor (e.g. gene-CNA) association map, and thus can not be applicable to identify master predictors. In addition, *Yang et al.* (2013) proposed a sparse multivariate regression model with low-rank representation to account for confounding factors (LORS). Our smFARM may be regarded as being equivalent to LORS, by setting our $ZB^T = L$, where L is a low rank matrix with rank K. Unlike LORS, which employs singular value decomposition (SVD) to construct a low-rank representation of L, we use a factor analysis model to directly and explicitly decompose the L as factors Z and corresponding factor loadings B for better biological interpretations. Furthermore, LORS assumes an isotropic noise $\epsilon \sim \text{MVN}(0, \sigma^2 I)$, whereas our smFARM assumes a unique variance structure with $\epsilon \sim \text{MVN}(0, \Psi)$. Apart from these modeling differences, our smFARM provides a framework that allows for more relevant and detailed interpretation about the residuals. In contrast, the interpretation for the L matrix does not seem to be straightforward.

2.6 Simulation Studies

2.6.1 Simulation Setup

We conduct three simulation experiments to assess the performance of the proposed model and optimization method. To specify simulation settings, we mimic a microarray data with $N = 200$ subjects, $Q = 400$ gene expressions and $P = 400$ variables of copy number alterations (CNAs). For each simulation, we consider a

specific association map between genes and CNAs which is sparse with groups. The graphic presentation of each map is given in Figure 2.1. In simulation experiment I, we begin with a simple association map, in which 5 CNAs are master regulators (or hubs) shown in Figure 2.1(a). These master CNAs are strong and totally link to 114 genes, on average each regulating 20 to 30 gene expressions. The total number of nonzero associations in this map is 125. In simulation experiment II, we investigate a more complex association map consisting of 46 weak CNA master regulators out of 400 CNAs with each connecting with 1 to 8 genes. The total number of nonzero associations is 183; see Figure 2.1(b). This situation is more challenging as it contains more clusters of low degrees in comparison to Figure 2.1(a). It is expected that the L_1 norm penalty in favor of individual signal selection would perform better than the group penalty. Simulation experiment III concerns a more practical situation, where the topology of an association map may be neither group dominated nor individual dominated. We consider a map shown in Figure 2.1(c) which includes 5 strong master CNAI regulators, each influencing 24 to 37 genes, 5 weak master CNAI regulators, each influencing 3 to 7 genes, and 20 CNAIs linking to only 1 or 2 genes. The total number of nonzero associations is 192.

In the first two simulation experiments I and II, P categorical CNAs $x = (x_1, \ldots, x_P)^T$ are generated as predictors from $x_p \sim \text{Binomial}(2, 0.2) - 1$, with values -1, 0, or 1, representing copy number deletion, normal and amplification. In the third simulation study, continuous copy number alternation intervals (CNAIs) are generated to mimic the true predictor characteristics discussed in Section 2.7. Based on the real breast cancer data, we find that there exits the heterogeneity within CNAIs, characterized by certain chromosome-specific structures, occurring in the forms of both within-chromosome and between-chromosome differences. Here we assume that these P continuous CNAIs belong to 23 distinct chromosomes, with the number of CNAIs (i.e. $P_i, i = 1, \ldots, 23$) on the i-th chromosome proportional to the size of

that chromosome obtained from the real data. Within the i-th chromosome, any pair of CNAIs, say, CNAI_m and CNAI_n, is set to be positively correlated and such correlation decreases when their genetic distance increases according to $0.9^{|m-n|/2}$ for $m,\ n = 1, \ldots, P_i$. If two CNAIs come from different chromosomes, a much weaker correlation is randomly drawn from $\{0.25, 0.25^2, \ldots, 0.25^{23}\}$ together with a randomly generated positive or negative sign. Finally we compute the nearest positive definite symmetric matrix Ξ based on the above correlations using the algorithm in *Higham* (1988), and P continuous CNAs are generated from $x \sim \text{MVN}_P(0, \Xi)$.

To specify the $Q \times P$ association map of $\Theta = \{\theta_{qp}\}$, we first specify a sparse indicator matrix $\Delta = \{\delta_{qp}\}$ which defines the connectivity in a genetic association mapping between Q genes and P CNAs. If $\delta_{qp} = 1$, we generate θ_{qp} from $\text{Unif}([-5, -1] \bigcup [1, 5])$; otherwise, $\theta_{qp} = 0$. To specify the $Q \times K$ factor loadings matrix B, we start with an initial matrix $B^* = \{b^*_{qk}\}$, with $b^*_{qk} \overset{i.i.d.}{\sim} \text{Unif}([0, \tau])$ and τ is a given positive constant. Then specify matrix B as of the form $B = UV^{\frac{1}{2}}$, where V is a diagonal matrix with diagonal entries being the eigenvalues of $B^* B^{*T}$, and the column vectors of U are the orthonormal eigenvectors of $B^* B^{*T}$. In other words, matrix B is specified by an orthogonal rotation of the initial matrix B^*. Note that the factor loadings have an "indeterminacy" problem, which means both B and BT give rise to the same covariance matrix $\Sigma = BB^T + \Psi$, where T is an arbitrary orthogonal matrix. To ensure a unique solution, we impose a constraint on B that $B^T B$ is a diagonal matrix (*Anderson and Rubin*, 1956). Our procedure of generating the values of factor loadings for matrix B accounts for such constraints. Given Θ and B, for each subject, we generate K latent factors $z = (z_1, \ldots, z_K)^T$ by $z_k \sim \text{Normal}(0, 1)$ and Q measurement errors $\epsilon = (\epsilon_1, \ldots, \epsilon_Q)^T \sim \text{MVN}_Q(0, \Psi)$, where the uniqueness Ψ is set as $\Psi = \sigma^2 I_Q$ in the simulation studies. Recall that τ and σ^2 are two constants that control the size of communality and that of uniqueness, respectively. The choice of both τ and σ^2 is based on a pre-specified scale of signal-to-noise ratio, according to SNR_1 of regression

mean effects and SNR_2 of latent factor's effects, namely $\mathrm{SNR}_1 = \mathrm{avg}\left[\frac{\mathrm{diag}(\mathrm{Cov}(\Theta x))}{\mathrm{diag}(\mathrm{Cov}(\epsilon))}\right]$ and $\mathrm{SNR}_2 = \mathrm{avg}\left[\frac{\mathrm{diag}(\mathrm{Cov}(Bz))}{\mathrm{diag}(\mathrm{Cov}(\varepsilon))}\right]$, respectively. Finally, Q gene expressions $y = (y_1, \ldots, y_Q)^T$ are generated from model (2.3) by $y|x, z \sim \mathrm{MVN}_Q(\Theta x + Bz, \Psi)$. Hereafter, a dataset of N i.i.d. (y, x) pairs is generated for each simulation round.

For convenience, responses and predictors are all centered to mean zero and the prior knowledge matrix $C = \{C_{qp}\}$ is set as all entries are 1, namely all predictors are subject to shrinkage. Our primary evaluation criterion is the total number of false discoveries, TF = FP + FN, where FP and FN are the respective numbers of false positives and false negatives. Here, a "positive" (or a "negative") refers to a nonzero (or a zero) entry of Θ. Following *Fan et al.* (2009), additional criteria used in the evaluation include sensitivity (Sen), and Matthews correlation coefficient (MCC) score defined respectively, by $\mathrm{Sen} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, and $\mathrm{MCC} = \frac{(\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$.

To assess the performance of our smFARM, we mainly compare it with remMap by varying SNR_1, SNR_2 and K. It is worth noting that *Peng et al.* (2010)'s remMap approach, which is established for the classic multivariate regression models (i.e. $K_{\mathrm{true}} = 0$), has been compared with two popular existing methods, single LASSO penalty (i.e. $\lambda_2 = 0$) and Q separate individual LASSO regressions, and its superiority has been showed in the paper (*Peng et al.*, 2010). So the comparisons to the latter two methods are not reported in our comparison. In the first part of Simulation I (denoted as I.1), we simulate data without any latent factors and aim to show the consistency between the remMap and our proposed smFARM method. Then in the rest of simulation experiments, the second part of Simulation I (denoted as I.2), II and III, we set the true number of latent factors as $K_{\mathrm{true}} = 2$, and focus on comparing three scenarios with $K = 0$ (i.e. remMap), $K = K_{\mathrm{true}}$ (i.e. 2), and $K = K_{\mathrm{BIC}}$. The BIC criterion is given in (2.11) over the range of 0 to 4, in which $tr[\widehat{H_Z}]$ is used as the estimated degrees of freedom. And K_{BIC} is also compared with K_{CV} from cross validation criterion. The tuning parameters (λ_1, λ_2) are determined through 5-fold

cross validation. And a total of 50 independently replicated datasets is used in the evaluation of our method. Results of method comparisons are summarized in Table 2.1.

2.6.2 Findings from Simulation Studies

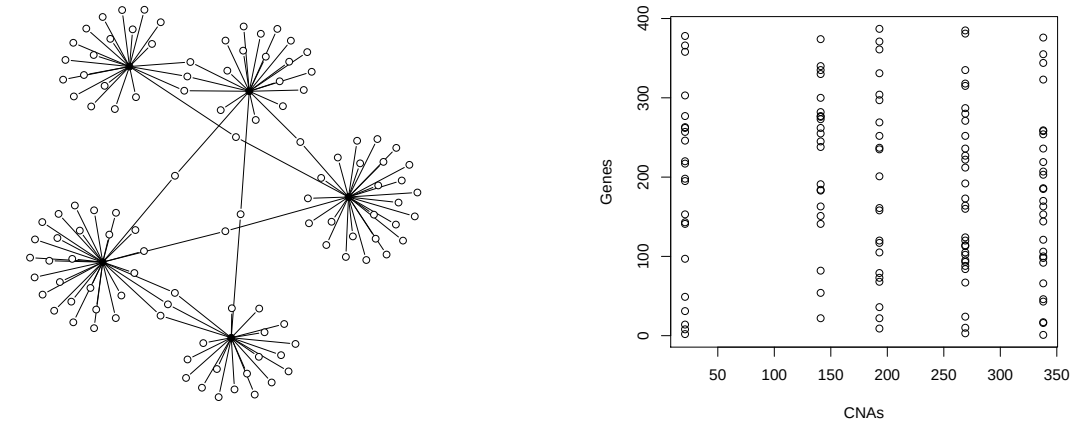
Let us first focus on simulation study I, including two cases I.1 and I.2, the corresponding numerical results are reported in the top part of Table 2.1. In Simulation I.1, when the true model contains no latent factors, subject to rounding errors, our smFARM and the remMap perform equally well in terms of MCC. With no surprise, we find that, in both smFARM and remMap, larger SNR_1 leads to better performance in terms of lower TF, higher Sensitivity and higher MCC through the comparison between SNR=1:0:3 and SNR=1:0:5. This finding is repeated by the comparison between SNR=1:1:3 and SNR=1:1:5 with $K_{\text{true}} = 2$ in Simulation I.2. When the ratio of SNR_1 to SNR_2 is fixed at 1:1, smaller variation in the measurement errors (i.e. larger SNR_1) will lead to better performances. Moreover, an encouraging finding in Simulation I.2 is that, comparing our method accounting for the latent factors to the remMap ignoring the latent factors, the smFARM approach is clearly more effective to identify true signals than the remMap when $K_{\text{true}} \neq 0$. In addition, both simulation studies I.1 and I.2 show us that our BIC criterion works quite well to determine the true number of latent factors. With fixed SNR_1, when comparing $(\text{SNR}, K_{\text{true}}) = (1{:}0{:}3, 0)$ in Simulation I.1 with $(\text{SNR}, K_{\text{true}}) = (1{:}1{:}3, 2)$ in Simulation I.2, and $(\text{SNR}, K_{\text{true}}) = (1{:}0{:}5, 0)$ in Simulation I.1 with $(\text{SNR}, K_{\text{true}}) = (1{:}1{:}5, 2)$ in Simulation I.2, we obtain very similar results of our smFARM when the latent factors are accounted for. These findings also suggest that SNR_2 has a strong influence on the performance of the remMap when the dependency of latent factors is ignored in the analysis.

It is interesting to note that results of group selection in simulation study I are

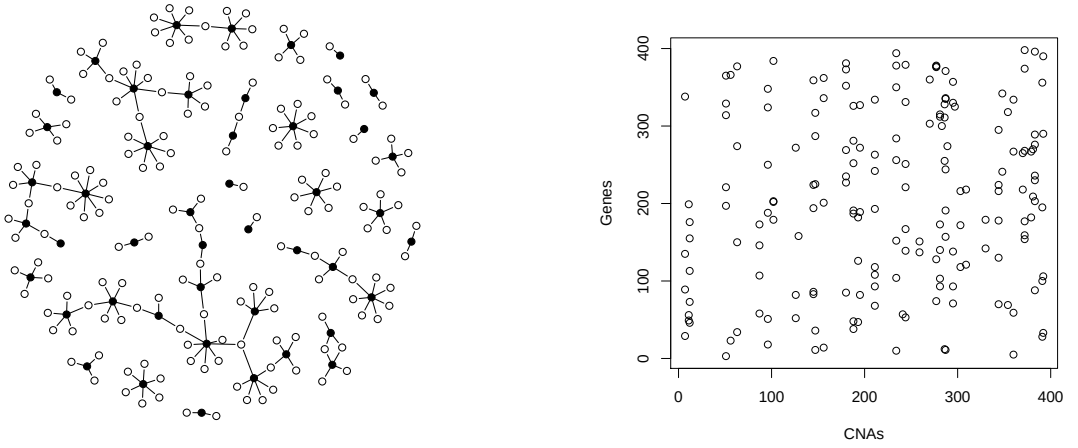
rather stable and accurate across the four cases in the top part of Table 2.1. This is probably because identifying clusters in these settings is not hard due to group-dominant topology designed in the association maps (also see Figure 2.1(a)). In other words, relative to the L_1 penalty, the L_2 penalty is more effective to remove irrelevant groups or clusters.

The results given in the middle part and the bottom part of Tables 2.1 concern simulation studies II and III. Once again these results show that the proposed smFARM performs very well in all key aspects of regulator detection, group selection and latent factor identification. Taking Simulation II as an example, when evaluating the performance of smFARM in the case SNR=1:3:5 among K=0, 2, 3, smFARM selects the true number of latent factors (i.e. 2) with 100% success rate by both K_{BIC} and K_{CV}. Moreover, both regulator selection and group selection show us that smFARM achieves the highest sensitivity and MCC as well as the lowest total false rate. From the comparison between $(\mathrm{SNR}, K_{\mathrm{true}}) = (1{:}0{:}5, 0)$ and $(\mathrm{SNR}, K_{\mathrm{true}}) = (1{:}3{:}5, 2)$, it is evident that when properly adjusting the latent factors, SNR_2 will have little influence on the reconstruction of the association map. Once again this summary implies that it is important to account for unobserved factors in the association analysis of high-throughput array data, and failing to adjust for such underlying heterogeneity will incur the loss of statistical power in the reconstruction of association map. In addition, all the above conclusions have repeated consistently in the more realistic simulation study III with continuous predictors. To sum up, our proposed method has demonstrated clearly as being a very effective tool to achieve desirable statistical power by accounting for latent factors in the regulatory map reconstruction with high-dimensional complex data.

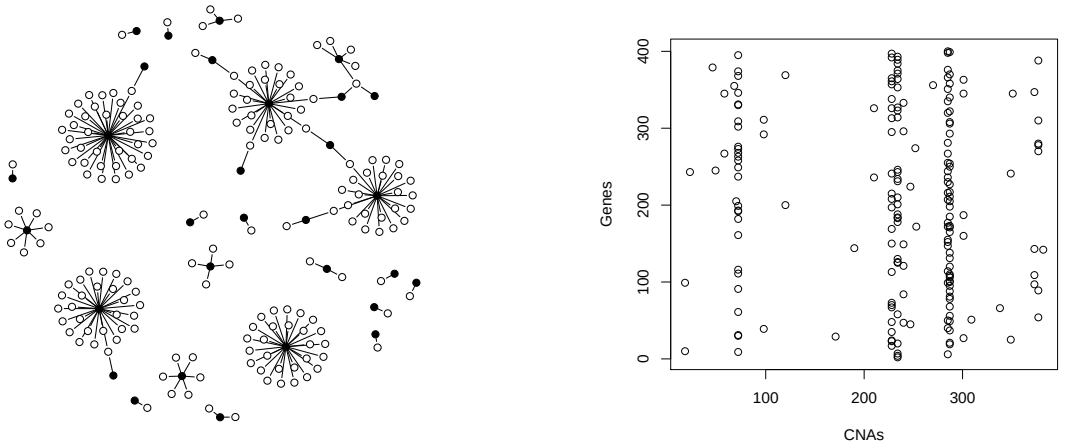
Figure 2.1: True association maps of Θ (connectivity vs. heatmap) for Simulation I, II, and III. (**LHS**: connectivity maps of Θ between genes (white) and biomarkers (black); **RHS**: corresponding heatmap of Θ.)



(a) Simulation setting I



(b) Simulation setting II



(c) Simulation setting III

Table 2.1: Results of Simulation I to Simulation III: Impact of different number of latent factors K and different SNR levels on Regulator Selection and Group Selection

			Regulator Selection			Group Selection				
SNR	K_{true}	Method	TF	Sen	MCC	TF	Sen	MCC	$K_{\text{BIC}}(\%)$	$K_{\text{CV}}(\%)$
Simulation I.1										
1:0:3	0	$\text{smFARM}_{\text{BIC}}$	18.90(6.02)	0.89(0.04)	0.92(0.02)	0.06(0.24)	1(0)	0.99(0.02)	0 (100%)	0 (100%)
		remMap	21.88(6.61)	0.93(0.02)	0.92(0.02)	0.02(0.14)	1(0)	1(0.01)		
1:0:5	0	$\text{smFARM}_{\text{BIC}}$	27.24(3.51)	0.81(0.03)	0.88(0.01)	0(0)	1(0)	1(0)	0 (100%)	0 (100%)
		remMap	34.10(5.17)	0.88(0.03)	0.87(0.02)	0(0)	1(0)	1(0)		
Simulation I.2										
1:1:3	2	$\text{smFARM}_{\text{BIC}}$	18.24(3.46)	0.87(0.03)	0.92(0.01)	0(0)	1(0)	1(0)	2 (100%)	2 (100%)
		remMap	25.68(11.32)	0.83(0.04)	0.89(0.04)	0.02(0.14)	1(0)	1(0.01)		
1:1:5	2	$\text{smFARM}_{\text{BIC}}$	28.51(4.26)	0.80(0.03)	0.88(0.02)	0(0)	1(0)	1(0)	2 (100%)	2 (100%)
		remMap	33.40(4.92)	0.76(0.04)	0.86(0.02)	0(0)	1(0)	1(0)		
Simulation II										
1:0:5	0	$\text{smFARM}_{\text{BIC}}$	63.36(6.42)	0.66(0.04)	0.81(0.02)	5.70(1.64)	0.88(0.04)	0.93(0.02)	0 (100%)	0 (100%)
1:1:5	2	$\text{smFARM}_{\text{BIC}}$	65.67(6.17)	0.65(0.04)	0.80(0.02)	5.94(1.66)	0.87(0.04)	0.93(0.02)	2 (100%)	2 (100%)
		$\text{smFARM}_{K=0}$	74.41(6.92)	0.60(0.04)	0.77(0.02)	7.80(1.70)	0.83(0.04)	0.90(0.02)		
		$\text{smFARM}_{K=3}$	94.74(11.12)	0.49(0.06)	0.69(0.04)	16.87(4.22)	0.63(0.09)	0.78(0.06)		
1:3:5	2	$\text{smFARM}_{\text{BIC}}$	66.32(6.15)	0.64(0.04)	0.80(0.02)	6.24(1.80)	0.87(0.04)	0.92(0.02)	2 (100%)	2 (100%)
		$\text{smFARM}_{K=0}$	107.86(13.85)	0.43(0.06)	0.64(0.05)	15.74(4.99)	0.66(0.11)	0.79(0.07)		
		$\text{smFARM}_{K=3}$	89.84(12.32)	0.52(0.07)	0.71(0.05)	13.53(4.16)	0.71(0.09)	0.82(0.06)		
Simulation III										
1:3:5	2	$\text{smFARM}_{\text{BIC}}$	48.89(11.54)	0.82(0.05)	0.87(0.03)	10.89(2.53)	0.66(0.06)	0.79(0.05)	2 (94 %)	2 (92 %)
		$\text{smFARM}_{K=0}$	79.80(16.76)	0.77(0.02)	0.79(0.04)	12.10(1.25)	0.62(0.04)	0.76(0.03)		
		$\text{smFARM}_{K=3}$	85.39(15.49)	0.64(0.09)	0.75(0.05)	13.00(4.17)	0.65(0.06)	0.75(0.07)		
		remMap	87.46(20.67)	0.79(0.03)	0.77(0.05)	12.46(1.35)	0.62(0.05)	0.75(0.03)		

Note: when tuning the number of latent factors via BIC or cross validation, we list the percentage of selecting K_{true} on 50 replicates. And for each Total False (TF), Sensitivity (Sen), or Matthews correlation coefficient (MCC) measurement, we also report mean values together with their standard errors. $\text{smFARM}_{K=K_0}$ represents fitting the smFARM on a given number of latent factors K_0.

2.7 Application

We now apply the proposed smFARM to analyze a breast cancer dataset, which has been previously analyzed in *Peng et al.* (2010) using remMap without accounting for latent factors. The data is measured by CGH arrays and cDNA expression arrays over 172 breast cancer tumor samples, and then is preprocessed in the same way as was done in *Peng et al.* (2010). Briefly, based on the CGH array data of about 17K genes, the genome is divided into 384 copy number alteration intervals (CNAIs). The DNA copy number status of these CNAIs is treated as predictors in the analysis. In addition, RNA transcripts of 654 breast cancer related genes, a union set of 7 published breast cancer gene lists, are treated as responses in the analysis.

Our primary goal is to reconstruct a regulatory association map between copy number alterations and RNA expressions, adjusting for tumor subtypes and potential latent factors. The estimated loadings from the factor model can provide information for additional genetic or non-genetic features left in the residuals. For each pair of CNAI and RNA transcript, it can be classified as a linked (or cis-) pair, if the transcript falls in the genome region of the CNAI; or otherwise an unlinked (trans-) pair. Based on this definition, there are totally 519 linked pairs identified. In addition, tumor subtypes could be important confounders affecting the construction of genetic regulatory map in array CGH analysis. Thus, the 172 samples are divided into 5 subtypes based on their expression profiles (*Peng et al.*, 2010). In our analysis, we include four subtype dummy predictors to adjust for the subtype-specific effects. Our analysis is based on the following mFARM:

$$Y_{\text{RNA}} = X_{\text{CNAI}}\Theta^T + G_{\text{subtype}}\Pi^T + ZB^T + E, \tag{2.12}$$

where Y_{RNA} is a 172×654 RNA expression matrix, X_{CNAI} is a 172×384 CNAIs output matrix and G_{subtype} is a 172×4 tumor subtype indicator matrix, Θ and Π

are two regression coefficient matrices with respect to CNAIs and tumor subtype. In addition, $P(= 654)$ responses Y_{RNA} and $Q(= 389)$ predictors (X_{CNAI}, G_{subtype}) are centered to zero.

One of our objectives is to detect potential trans-regulations, by adjusting for all 519 cis pairs. This can be set up by letting $C_{qp} = 0$ if q-th CNAI links to p-th gene; and $C_{qp} = 1$, otherwise. In this way, there will be no shrinkage imposed on the coefficients of cis pairs.

To evaluate the performance of the proposed regularization procedure in terms of selection stability, we generate 100 bootstrap samples from the original dataset. For each bootstrap sample, we select tuning parameters (λ_1, λ_2) using 10-fold cross validation on a 25×30 grid. As pointed above, the mechanism of biological pathways underlying breast cancer is notoriously complex, so that the associated genetic heterogeneity may induce a large number of genetic and/or non-genetic latent factors. For example, *Lucas et al.* (2010) obtained 56 latent factors in a breast tumor study. In practice, however, researchers would like to focus on several leading latent factors for more meaningful biological interpretations. For instance, *Carvalho et al.* (2008) controlled the number of latent factors by an upper bound $K \leq 10$ in a breast cancer hormonal study. In this analysis, we select the number of latent factors K over a range of 0 to 10 using the BIC in (2.11) with degrees of freedom $\widehat{df}(K) = tr[\widehat{H_Z}] + K$, where $tr[\widehat{H_Z}]$ and K quantify the respective complexity of the mean model and factor analysis model. Under this BIC criterion, among 100 bootstrap samples, the number of latent factors K=5 has been selected with the highest frequency 50%, then K=6 with 28% and K=4 with 22%. The results of stability for the regulatory relationships are listed in Table 2.2. Note that, any trans-regulations with 10% selection rate or lower are not reported in Table 2.2.

As seen from Table 2.2, two CNAIs on 17q12 are associated with multiple unlinked genes. Specifically, compared to the results in *Peng et al.* (2010) using remMap, we

Table 2.2: Association map detection frequencies over 100 bootstrap samples.

Trans-group Selection		Transcripts being trans-regulated			
CNAI	Nucleotide position(bp)	Clone ID	Gene Symbol	Cytoband	Freq.(%)
CNAI1	17q12, 34811630-34811630	68400	BM455010	17	65
		418240	LOC90110	17q21.2	43
		159608	-	-	18
		270535	BM466581	19	15
		756931	S100A1	1q21	14
CNAI2	17q12, 34944071-35154416	159608	BM455010	17	19
		854899	DUSP6	12q22-q23	18
		1337808	DUSP6	12q22-q23	16
		725321	ESR1	6q25.1	14
		503602	CAMK2N1	1p36.12	12

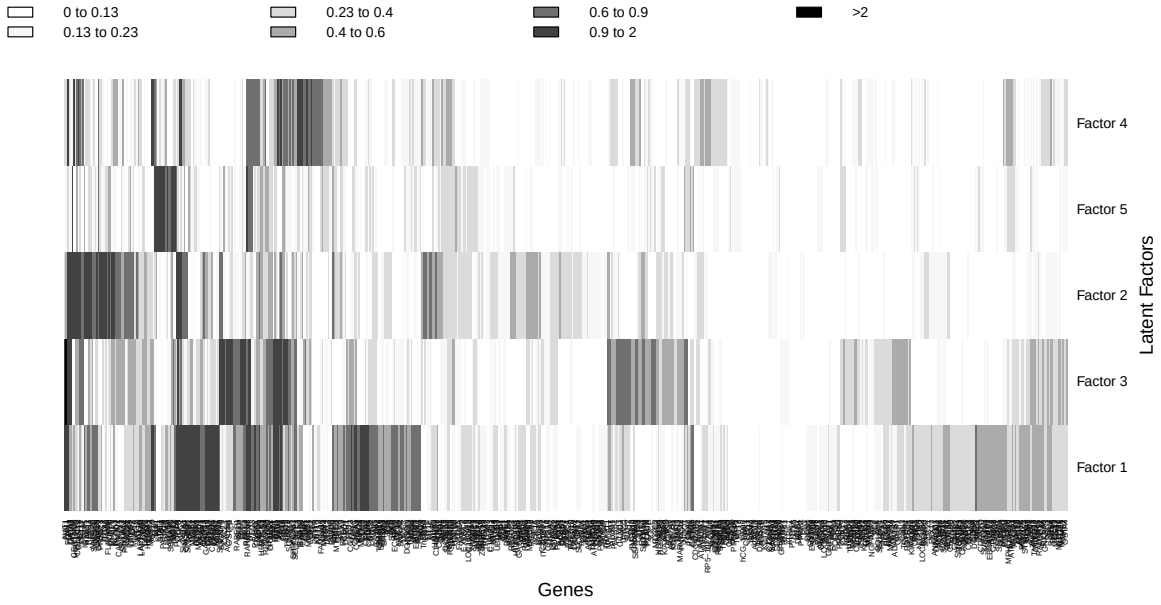
Note: gene BM455010 does not fall into the CNAI1 or CNAI2, but sits very closely to them. This probably should be treated as a cis regulation.

have detected additional regulation relationships between CNAI2 and genes of ESR1, DUSP6, and CAMK2N1. These new findings are highly biologically relevant and quite intriguing. CNAI2 is part of the ERBB2 amplification region, and the amplification of ERBB2 is the key characteristics of the "HER+" subtype breast tumors. On the other hand, ESR1 encodes an estrogen receptor, which is over-expressed in around 70% of breast cancer cases, referred to as the "ER+" subtype. Our findings here suggest the potential interplay between these two important oncogenes of breast cancer—ERBB2 and ESR1. Discoveries given in our analysis helps to shed light on better understanding of targeted molecular regions and mechanism for breast cancer.

Besides the above findings of trans-regulation relationship, the results from smFARM also suggest that CNAIs together with other covariates only explain a limited proportion of the variation of gene expressions, and the latent factors serve as a useful source for additional knowledge generation. Here we refit smFARM on this breast cancer data with K=5, which has the highest selection rate from 100 bootstrap samples. To make sense of biological interpretation on these 5 latent factors, we apply the varimax orthogonal rotation method (*Kaiser*, 1958) on the estimated loadings $\hat{B}$ to achieve

sparse relationships between the latent factors and the regulated genes. The heatmap of $\hat{B}$ after the varimax rotation is displayed in Figure 2.2, which provides us a graphic overview of the relative sparsity and block skeleton of these latent factors, as well as the biological crosstalk in terms of genes linked to multiple latent factors.

Figure 2.2: Breast cancer hormonal pathways are displayed by the heatmap of gene-factor loadings $|\hat{B}_{q,k}|$ after varimax rotation from the fitted smFARM model for the 654 selected genes and 5 factors.



When gene sets share genes, examination of how they overlap can highlight common processes, pathways, and underlying biological themes. To further unveil the potential biological functions of these 5 latent factors, we then employ a web-based statistic tool (`http://www.broadinstitute.org/gsea/msigdb/help_annotations.jsp#overlap`) to perform enrichment analysis for specific lists of probe sets. The assessment of enrichment is based on an overlap statistic (e.g. *Chang and Nevins* (2006), *Mootha et al.* (2003), *Subramanian et al.* (2005)), such as Fisher exact test in a two-way annotation contingency table, which evaluates its overlap with a user provided gene set, and an estimate of the statistical significance, with certain chosen annotations. The sources of annotations considered in our analysis include a combi-

nation of canonical pathways (CP), KEGG gene sets and Gene Ontology (GO) gene sets.

We then annotate these breast cancer related genes through the identified latent factors. By testing whether a set of "top" genes within each latent factor (i.e. by top genes, we mean those genes that have strong loadings on the latent factors with cutoffs of 25% and 75% quantiles) are significantly enriched in the above CP, GO and KEGG loci, we can learn how these top gene components overlay multiple aspects of known biological activities. The corresponding results are presented in Table 2.3.

We find that, when adjusting the CNAIs effect, latent factors 1, 3 and 5 are significantly enriched in some known genetic pathways with $p < 1.00e^{-5}$. Since these 654 breast cancer related genes are preselected from the published breast cancer gene lists, including many cell cycle-regulated genes, it is not surprising to observe that some genes from the enrichment analysis show significant over-representation of cell cycle characteristics. It is of great interest to note that we have identified two gene sets that do not belong to cell cycle pathways: *chromosome*(GO:0005694) gene set and *FOXM1* pathway. It is suggested that malfunction of genes in *chromosome* (GO:0005694) may result in uncontrolled cell proliferation and cancer development. *Mosca et al.* (2010) reported that there are 71 genes from *chromosome* (GO:0005694) to be altered in breast cancer cells. From our enrichment analysis, we have detected 17 genes in overlap with *chromosome* (GO:0005694), 11 of which are from the pre-specified list of 71 breast cancer related genes. *FOXM1* transcription factor network in Forkhead signaling pathways also includes some important cancer genes, such as BRCA2, ESR1 and FOXM1. Mutations in gene BRCA2, which is involved in DNA repair pathway, causes about half of the cases of early-onset breast cancer (*Sharan et al.*, 1997); gene ESR1, as mentioned above has found to be frequently amplified in breast cancer (*Holst et al.*, 2007); and gene FoxM1 is tightly regulated during the cell cycle in terms of expression and transcriptional activity, as well as the development

Table 2.3: Summary of biological terms characterizing factors adjusting for CNAIs effect ($p < 1.00e^{-5}$).

	Type	Pathway	No. Genes in Gene Set	No. Genes in Overlap	p-value
Factor 1	GO	CELL CYCLE PROCESS	193	27	$9.65e^{-11}$
	GO	MITOTIC CELL CYCLE	153	24	$9.76e^{-11}$
	GO	CELL CYCLE GO:0007049	315	33	$1.85e^{-09}$
	GO	CELL CYCLE PHASE	170	23	$4.76e^{-09}$
	GO	M PHASE	114	17	$1.20e^{-07}$
	CP	REACTOME CELL CYCLE MITOTIC	325	30	$1.82e^{-07}$
	GO	CHROMOSOME	124	17	$4.15e^{-07}$
	KEGG	KEGG CELL CYCLE	128	15	$8.85e^{-07}$
	GO	M PHASE OF MITOTIC CELL CYCLE	85	12	$1.53e^{-06}$
	GO	MITOSIS	82	13	$1.84e^{-06}$
Factor 3	GO	MITOTIC CELL CYCLE	153	18	$3.31e^{-08}$
	GO	CELL CYCLE PROCESS	193	20	$4.93e^{-08}$
	CP	PID FOXM1PATHWAY	40	9	$3.81e^{-07}$
	GO	M PHASE	114	14	$6.77e^{-07}$
	GO	CELL CYCLE GO:0007049	315	24	$7.73e^{-07}$
	GO	CELL CYCLE PHASE	170	17	$8.48e^{-07}$
	GO	M PHASE OF MITOTIC CELL CYCLE	85	12	$9.51e^{-07}$
	GO	MITOSIS	82	11	$4.53e^{-06}$
Factor 5	GO	CELL CYCLE PROCESS	193	16	$6.43e^{-08}$
	GO	CELL CYCLE GO:0007049	315	19	$5.62e^{-07}$
	GO	MITOTIC CELL CYCLE	153	13	$8.74e^{-07}$
	CP	REACTOME CELL CYCLE MITOTIC	325	19	$9.03e^{-07}$
	CP	PID PLK1 PATHWAY	46	7	$6.99e^{-06}$

Note: Top overlaps with cutoff $p < 1.00e^{-5}$ are chosen because the results from another web-based statistical tool Gather (see *Chang and Nevins* (2006), `http://gather.genome.duke.edu/`) are highly matched these top gene sets from Fisher exact test when $p < 1.00e^{-5}$.

and progression of many malignancies, including breast cancer (*Ahmad et al.*, 2010).

In summary, we have found that latent factors 1, 3 and 5 are closely tied with genes involved in cell division and duplication in breast cancer. In contrast, no pathways have been found to be significantly associated with factors 2 and 4, implying that these two factors are possibly tied with unknown non-genetic causes, such as environmental conditions, batch effects and population heterogeneity.

2.8 Discussion

We developed a new methodology, sparse multivariate factor analysis regression model, to reconstruct a sparse genetic association map. The proposed smFARM extends the classic multivariate regression model to allow a low-dimensional set of latent factors to account for the dependence among response variables instead of assuming residuals being i.i.d. noise. Through gene enrichment analysis, our method helps to explore additional "nuggets" in the residuals that can enhance the understanding of the underlying data generation mechanism. We also developed an effective and flexible EM-GCD algorithm to obtain regularized estimation and variable selection in the smFARM.

We have shown that by accounting for a suitable number of latent factors, the proposed smFARM can effectively identify response-predictor associations from high dimensional data with improved sensitivity and accuracy. The numerical results have indicated that the proposed smFARM works well to derive not only the underlying sparse association relationship but also the number of latent factors. In contrast, if variations explained by these latent factors are not modeled properly, the resulting association map given by the existing methods such as remMap is of low power. Furthermore, the real breast cancer data example has also shown that our proposed smFARM provides richer and biologically relevant discoveries to facilitate transcriptomic analyses. Not only can we construct the sparse genetic association map between

CNAIs and gene expressions, but also perform in-depth gene enrichment analysis for the latent factors, which helps us to understand and interpret the residual features tied with either genetic or non-genetic regulations and mechanisms. This latter is the new contribution from our proposed method.

To our knowledge, there are some other methods that can characterize the variability in the gene expressions such as singular value decomposition (SVD) or principle component analysis (PCA). There is a direct relationship between PCA and SVD in the case where principal components are calculated from the covariance (*Wall et al.*, 2003). Furthermore, the essential difference between SVD/PCA and factor analysis lies whether or not a covariance model is used for the residuals. Refer to *Schneeweiss and Mathes* (1995), *Tipping and Bishop* (1999) and *Van Wieringen and Van De Wiel* (2011) for more details. We find that unlike PCA/SVD using superficial labeling such as "eigengenes", "supergenes", or "meta-genes" without clear biological entity (*Alter et al.*, 2000), the latent factors can provide meaningful and relevant biological interpretation in the reconstruction of association map, which is appealing in practice.

Although we have focused on a combination of L_1 and L_2 penalties, some other penalties, such as adaptive LASSO and adaptive group LASSO may be established in our method with minor modifications. Moreover, if the covariance matrix Σ of the response variables is sparse, we may further consider regularizing the factor loadings B and simultaneously penalizing coefficient matrix Θ and loading matrix B. Besides the gene-CNA association analysis illustrated in this chapter, our proposed method may be applied in a broad range of problems. For instance, it may be applied to systematically explore the relationship between gene expression levels and genotypes as to, for example, whether a gene is differentially expressed with different genotypes (or alleles) at a specific locus. The loci that are associated with gene expression levels are known as expression quantitative loci (eQTL). For a given gene, an eQTL data analysis aims to identify genetic loci or single nucleotide polymorphisms (SNPs) that

are linked or associated with expression levels of a common gene. Moreover, in eQTL analysis, SNPs may be naturally grouped according to their functionality or biological pathways based on some prior knowledge. When we are interested in associations of multiple SNPs simultaneously within a biological pathway, incorporating genetic or non-genetic latent factors would help us to achieve a more powerful and richer analysis, leading to better understanding of the underlying biological mechanisms.

CHAPTER III

Sparse structural factor equation model and its applications to the reconstruction of genetic regulatory networks

Directed acyclic graph (DAG) or Bayesian network has been widely utilized in the literature to represent causal relationships among genetic variants. A DAG describes the causal dependency structure among nodes in a network. When there exists a natural ordering among nodes, the objective of establishing a DAG is equivalent to estimating the network structure, which can be challenging due to the fact that some causal relationships may be obscured by unobserved shared factors. These latent factors may result in undirected edges among nodes, and thus the resulting network contains both directed and undirected edges, the so-called mixed graph. In this chapter, we mainly focus on a special class of mixed graphs – acyclic directed mixed graphs (ADMGs), which includes two subgraphs. One subgraph is a DAG with all directed edges, and the other is an undirected graph with all undirected edges. Structural equation model (SEM) is an appealing tool to formulate causal relationships in a DAG, but cannot be directly used if the existence of undirected edges is known. Hence, I propose a new graphical modeling approach, called *the sparse structural factor equation model (SFEM)*, in which I use the structural equation model for DAGs, while

accounting for potential latent factors using a factor analysis model. Utilizing latent factors, I hope to (i) identify and remove undirected edges induced by unobserved shared factors (i.e. common latent factors, such as unmeasured environmental factors); (ii) adjust for undirected edges in an ADMG that are annotated by the available published findings, using the means of node augmentation that enables us to convert an ADMG into a DAG. In this way, I yield a simpler and more interpretable causal network. The proposed SFEM is evaluated and compared to the existing methods (e.g. PC-algorithm) through extensive simulation studies, as well as real, proteomic data for the construction of human cell signaling pathways.

3.1 Introduction

Biological molecules in living organisms such as mRNAs and proteins do not function in isolation; rather they work together and interact with each other via an comprehensive functional network. Reconstruction of gene regulatory network (GRN) using gene expression data is of great importance for understanding gene functions and cellular dynamics in system biology as it pertains to the crucial knowledge of regulatory mechanisms in biological processes. Physical gene-gene interactions among individual genes can be experimentally derived by identifying transcription factors and their regulatory target genes, so can protein-protein interactions. However, such an experimental approach is time-consuming and labor intensive. The technological innovations in recent years allow gene expression levels to be measured for the whole genome simultaneously and across collections of related samples. These genome-wide expression data provide valuable information that can be fruitfully exploited to infer the network structure. In fact, a number of computational methods have been proposed in the literature to estimate gene networks from gene expression data.

Graphical model is currently a popular tool used for gene network inference. It has been useful to analyze and visualize conditional independence relationships among

variables of interest. Major components of a graphical model include nodes representing random variables and edges encoding relationships between the enclosing nodes. Based on whether edges have directions, graphical models have been classified into two types: a directed graphical model and an undirected graphical model. A directed graphical model (or Bayesian network) is a special type of a graphical model, whose dependence structure is represented by a directed acyclic graph (DAG). The utility of DAG for inferring causality has received much attention in the recent literature, in particular, its applications to the reconstruction of gene regulatory networks (*Friedman et al.*, 2000; *Segal et al.*, 2003; *Hartemink et al.*; *Peer et al.*, 2001).

Although DAGs have attractive properties in causal inference, when learning causal relationships in practice, sometimes not only the directed edges may be detected, but also undirected edges. And the resulting graph may lead to mixed graphs or even undirected graphs (*Anandkumar et al.*, 2013). Hence, we ought to consider a more general form of DAG, in which undirected edges are allowed. In this chapter, we consider a subclass of mixed graphs, named as acyclic directed mixed graphs (ADMGs), which are constructed by both DAGs and undirected graphs. Furthermore, we are interested in two different types of ADMGs. One is a DAG with numerous undirected edges, which could be caused by some unobserved common factors. For example, in a gene network analysis, latent factors may include genes that have not been included in the microarray, environmental factors, and latent population structure among the samples. Unfortunately, these factors or their effects could not be directly measured in the experiment and could influence all the observed gene expressions. And an exploratory analysis refers to the reconstruction of causal relationships while cleaning out the undirected edges induced by common latent factors. On the other hand, another type of ADMG is a DAG with a small proportion of undirected edges, and these undirected edges could be preselected from expert's empirical knowledge or learned directly from experiments. In this case, we do not assume that common

latent factors exist, and we call the detection of causal relationships by cleaning out a limited number of preselected undirected edges as confirmatory analysis. In addition, to our knowledge, the existence of undirected edges may distort causal relationships without being accounted for in the data analysis. Thus, it is critical to adjust for such undirected edges if exist, because if ignored, the analysis would yield either an excessive amount of false positives or reduced statistical power. To address this issue, we develop a new regularized estimation method to reconstruct DAGs with a known ordering.

Learning the dependence structure of a DAG from data presents a great challenge due to the fact that the number of candidate DAGs can grow super-exponentially along with the number of nodes (*Robinson*, 1973). The existing approaches for the DAG structure learning can be roughly categorized into three classes (*Schmidt*, 2010): search-and-score approaches, constraint-based approaches and hybrid approaches. A search-and-score approach attempts to learn a DAG structure by optimizing some criteria, such as the BIC or validation set likelihood, using either a search algorithm (*Lam and Bacchus*, 1994; *Heckerman et al.*, 1995) or Bayesian posterior distribution (*Friedman and Koller*, 2003; *Ellis and Wong*, 2008; *Zhou*, 2011). A constraint-based approach tries to prune a set of possible edges identified by conditional independence hypothesis tests, including the well-known Peter-Clark (PC) algorithm (*Spirtes et al.*, 2000), or by removing conditional dependencies that fall below a threshold (*Cheng et al.*, 2002). A hybrid method uses the constraint-based reasoning to prune a set of edges considered in the setting of a score-based method, which has been developed to improve computational efficiency (*Li and Yang*, 2004; *Tsamardinos et al.*, 2006). Recently, *Kalisch and Buhlmann* (2007) proposed a computationally efficient implementation for the PC algorithm to search sparse high-dimensional DAGs with polynomial complexity. However, the associated computational burden remains a big hurdle due to the size of the space of large DAGs.

A vast majority of the recent work has focused on the reconstruction of a sparse DAG through the penalized likelihood approach. In the special case where a topological ordering of the nodes is given, learning the structure of a DAG is equivalent to sparse estimation of modified Cholesky decomposition of a concentration matrix (i.e. the inverse of covariance matrix) (*Li and Yang*, 2005; *Huang et al.*, 2006; *Levina et al.*, 2008; *Shojaie and Michailidis*, 2010), which is computationally feasible. The information of node ordering is usually determined by a natural ordering of temporal observations, previous experiments and *a priori* knowledge (*Shojaie and Michailidis*, 2010). For example, when learning GRNs for microarray data, *a priori* knowledge of the node ordering could be obtained from the existing annotation software such as Cytoscape (*Lopes et al.*, 2010). If there is no established knowledge concerning the topological ordering, or each node is allowed to have more than one parental node, the penalized likelihood technique has proven to be computationally intricate (*Van de Geer and Bhlmann*, 2013). Combining a method of enforcing acyclicity with a block coordinate descent algorithm, *Fu and Zhou* (2013) and *Aragam and Zhou* (2014) developed some penalized methods for the estimation of DAG structures without *a priori* knowledge of the node ordering. However, from the above literature review, we find that no systematic work has been done in constructing sparse causal relationships under the framework of ADMG. Thus, it leads to the key interest of this research topic.

This chapter concerns a sparse estimation of DAGs using the proposed structural factor equation model, where accounting for latent factors is undertaken by the factor analysis model. In the presence of latent factors, learning DAG structure suffers the problem of parameter identifiability due to the fact that there may be multiple DAGs that equivalently explain the observed data (i.e. moral graph). In order to overcome this difficulty, one must restrict the set of possible graphical models. In this regard, I establish the needed criteria for parameter identifiability in the proposed structural

factor equation model.

The rest of this chapter is organized as follows. Section 3.2 introduces the proposed model on exploratory analysis and criteria for model identifiability, followed by the penalized estimation of DAGs based an EM-Coordinate-Descent (EM-CD) algorithm for numerical implementation in Section 3.3. Operating characteristics of the proposed method are examined on both simulated and real data in Sections 3.4 and 3.5, respectively. Section 3.6 presents the proposed model on confirmatory analysis. This chapter is concluded with discussion in Section 3.7, where the estimation of directed acyclic graphs with unknown ordering is discussed.

3.2 Structural factor equation model

3.2.1 Background and notation

Given a P-dimensional random vector $y = (y_1, \ldots, y_P)^T$ with a known partial ordering, we use a DAG $\mathcal{G} = (V, E)$ to describe causal relations among them. That is, each component y_i corresponds to one node in the DAG, with a directed edge between two nodes indicating a causal relation between them. Without loss of generality, we assume that y has been sorted according to its known ordering, which means a causal relation is only possible from variable y_j to variable y_i (i.e. $y_j \to y_i$) for $j < i$. The set of parental nodes of y_i is denoted by $pa(i) = \{j : j < i, y_j \to y_i\}$. Specifically, if for any $k < i$ and $k \notin pa(i)$, we have y_i is independent of y_k conditioning on $\{y_j\}_{j \in pa(i)}$.

To model causality among the components of y, we invoke a structural equation model (SEM) of the following form:

$$y_i = \sum_{j \in pa(i)} \theta_{ij} y_j + \epsilon_i, i = 1, \ldots, P, \tag{3.1}$$

where ϵ_i's are independent normal random errors with mean 0 and variance $\sigma_i^2 > 0$,

and ϵ_i is independent of y_i's parental nodes. Here $\Theta = \{\theta_{ij}\}$ is a $P \times P$ lower triangular matrix with zeros on the diagonal and is termed as *the weighted adjacency matrix* of a DAG $\mathcal{G}$.

Given a weighted adjacency matrix Θ of $\mathcal{G}$ and variance matrix of $\epsilon = (\epsilon_1, \ldots, \epsilon_P)^T$ $D = \text{diag}(\sigma_1^2, \ldots, \sigma_P^2)$, the SEM in (3.1) has the following first two moments of y. They are, $\mu = E(y) = 0$, and $\Sigma = \text{Cov}(y) = (I - \Theta)^{-T} D (I - \Theta)^{-1}$, and hence Σ, the covariance matrix of y, is uniquely determined by (Θ, D). Considering the concentration matrix $\Omega = \Sigma^{-1}$, we have a unique modified Cholesky decomposition given by $\Omega = (I - \Theta) D^{-1} (I - \Theta)^T$, where $I - \Theta$ known as the Cholesky factor that is a lower triangular matrix with the unit diagonals. This expression gives us an explicit connection between the DAG structure Θ and the concentration matrix Ω. It is worth pointing out that the above expression depends on the assumption of a known ordering of variables.

3.2.2 Structural factor equation model

To introduce latent factors into the SEM given in (3.1), we relax the assumptions that the error terms ϵ_i's are mutually independent. First, consider the following model with normally distributed dependent errors given in a matrix form:

$$y = \Theta y + \epsilon, \tag{3.2}$$

where normal variant ϵ has mean 0 and covariance W. We propose to model the covariance W by the following factor analysis model:

$$W = BB^T + \Psi, \tag{3.3}$$

where B is a $P \times K$ factor loading matrix for K ($\leq P$) latent factors and Ψ is a $P \times P$ diagonal matrix of uniqueness. Clearly, the mean of y is 0 while the covariance W of

error ϵ is determined by the factor loadings B and uniqueness Ψ.

Combining models (3.2) and (3.3) gives rise to the following structural factor equation model (SFEM):

$$y = \Theta y + Bz + e \tag{3.4}$$

where z is a K-variate vector of uncorrelated latent factors following multivariate normal distribution $\text{MVN}_K(0, I)$ and e is an error vector according to $\text{MVN}_P(0, \Psi)$ and is independent of z. Moreover, the first two moments of y are respectively, $\mu = 0$ and $\Sigma = (I - \Theta)^{-T}(BB^T + \Psi)(I - \Theta)^{-1}$. It is obvious that SFEM (3.4) reduces to the classical SEM with $K = 0$. From (3.4), we can also see that conditioning on the vector of K unobserved latent variables z, the vector of variables, y, satisfies the SEM for a DAG.

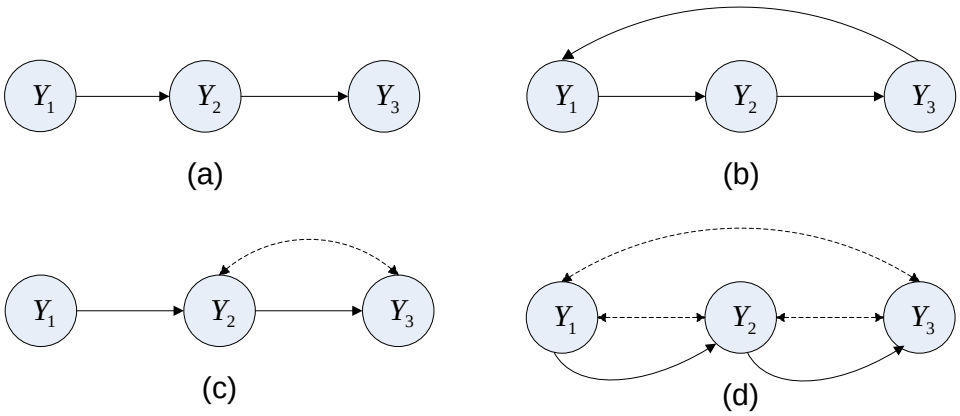
Note that the above analysis can be further formulated as an exploratory analysis (EA) or a confirmatory analysis (CA); in the latter, some entries of matrix B are set as zero in advance based on some established knowledge on an ADMG, whereas in the former, no *a priori* constraints are imposed and all entries of B will be determined by data. Hence, the proposed SFEM may be engaged with two different types of analyses in constructing a casual network.

3.2.3 Graphical representation of SFEM

As mentioned above, due to the potential influence of latent factors, we ought to consider a more general form of DAG, in which undirected edges are allowed. To proceed, we introduce some notations. Consider a special class of graphs, i.e. directed mixed graphs (DMGs), each of which includes both directed and undirected edges. A mixed graph is defined as $\mathcal{G} = (V, E, U)$, where V is a finite set of vertices and E, $U \subseteq V \times V$ are two disjoint sets of edges. The edges in E are directed or mono-directed; that is, $(i, j) \in E \Rightarrow (j, i) \notin E$, so we denote this kind of edge as $j \to i$. The edges in U are undirected or bi-directed; that is, $(i, j) \in E \Rightarrow (j, i) \in E$ and

vice versa, so we denote this type of edge by $i \leftrightarrow j$. The Figure 3.1 displays some examples of DMGs.

Figure 3.1: Four examples of directed mixed graphs. The graph in (b) is cyclic, while all others are acyclic. The solid line indicates an directed edge and the dashed line denotes an undirected edge.

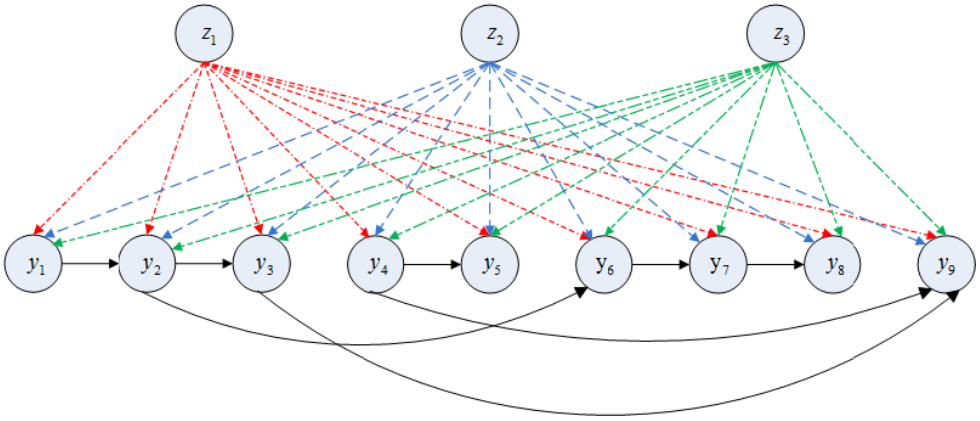


In this chapter, we focus on acyclic directed mixed graphs (ADMGs), a subclass of DMGs, which do not include directed self-loops (e.g. $(j, j) \notin E \cup U$). More specifically, an ADMG (V, E, U) consists of two subgraphs: one is a DAG (V, E) with all mono-directed edges, which may be modeled by a weighted adjacency matrix Θ, and the other is a subgraph of all undirected edges (V, U), which are determined by nonzero entries in the covariance matrix $W = BB^T + \Psi$ with $W_{ij} = W_{ji} \neq 0$ for $(i, j) \in U$ or $i = j$. As mentioned above, the proposed SFEM may be used in two different types of analyses when constructing a casual network. In the following sections, we focus primarily on the exploratory analysis (EA), while the confirmatory analysis (EA) is discussed in Section 3.6.

As mentioned above, EA is used when there is no *a priori* knowledge about matrices B and Θ. In the gene regulatory network study, common factors that relate to matrix B could be, for example, environmental variables, which are not measured but may alter gene expressions substantially. These factors could create many undirected edges, which impairs the power of identifying the underlying true causal relationships

among genes. For example, Figure 3.1 (d) shows that the directed chain network among nodes Y_1, Y_2 and Y_3 is fully masked by three undirected edges (dashed line). Thus, it is difficult to reconstruct a DAG without controlling the trigger of undirected edges. Figure 3.2 shows an example of ADMG modeled by SFEM: the vector of latent factors z_1, z_2 and z_3 in SFEM helps to account for the extra variations beyond what variables $y_1, \ldots, y_9$ can describe. In other words, the proposed SFEM may be used to construct a DAG among $y_1, \ldots, y_9$ conditional on three latent factors z_1, z_2 and z_3.

Figure 3.2: An example of SFEM for exploratory analysis: 3 common latent factors z_1, z_2 and z_3.



3.2.4 Parameter identifiability in SFEM

Parameter identifiability in SFEM for an exploratory analysis may be set up using the framework of factor models. That is, SFEM (3.4) may be rewritten as follows:

$$y = (I - \Theta)^{-1} B z + (I - \Theta)^{-1} e = \Gamma z + \delta, \tag{3.5}$$

where $\Gamma = (I - \Theta)^{-1} B$ and $\delta = (I - \Theta)^{-1} e$. Then, the resulting covariance matrix is $\Sigma = \Gamma\Gamma^T + \Sigma_\delta$ with $\Sigma_\delta = (I - \Theta)^{-1} \Psi (I - \Theta)^{-T}$. For model (3.5), we need to impose extra assumptions in order to identify both matrices Σ_δ and Γ.

- Identifiability condition (A) for the model (3.5): assume that $\Gamma^T \Sigma_\delta^{-1} \Gamma = B^T \Psi^{-1} B$

is diagonal with distinct entries arranged in a decreasing order;

- Identifiability condition (B) for matrix Σ_δ: assume that there exists a unique modified Cholesky decomposition of $\Sigma_\delta = (I - \Theta)^{-1}\Psi(I - \Theta)^{-T}$.

3.3 Penalized estimation

3.3.1 Formulation

When a natural ordering of the variables is available and the number of latent factors $K = 0$ (i.e. $B = 0$), the reconstruction of a DAG is equivalent to sparse estimation of the modified Cholesky decomposition of Σ_δ^{-1}. In this case, the identifiability condition (A) automatically holds. Several regularization approaches have been proposed to shrink elements in Θ to zero. See *Pourahmadi* (1999); *Wu and Pourahmadi* (2003); *Bickel and Levina* (2008); *Huang et al.* (2006); *Levina et al.* (2008), just name a few. More specifically, *Huang et al.* (2006) proposed adding an L_1 norm penalty on Θ to encourage zeros. *Levina et al.* (2008) proposed a banding procedure using a nested LASSO penalty. Recently, *Shojaie and Michailidis* (2010) employed the adaptive LASSO penalty to estimate the skeleton of DAG in the framework of SEMs and showed that this LASSO method is not sensitive to random permutations of the order of variables in y.

Given that our objective is to detect the sparse skeleton of DAG adjusting for latent factors, we propose the following penalized loss function:

$$\min_{\Theta} \frac{1}{2N} \sum_{n=1}^{N} (y_n - \Theta y_n)^T (BB^T + \Psi)^{-1} (y_n - \Theta y_n) + \lambda \sum_{i=1}^{P} \sum_{j=1}^{i-1} \xi_{ij} |c_{ij}\theta_{ij}|, \qquad (3.6)$$

where $y_n = (y_{n1}, \ldots, y_{nP})^T$ is the data from unit n, and λ is a nonnegative tuning parameter. The L_1 norm penalty term in the above loss function regularizes the sparsity in Θ.

If there is *a priori* knowledge about the causal relationship in y, such information may be incorporated into the optimization procedure. That is, define a pre-specified $P \times P$ matrix $C = \{c_{ij}\}$, whose (i, j)-th element is given by:

$$c_{ij} = \begin{cases} 1, \text{ if there is no prior information between } j \text{ and } i, \text{ when } j < i; \\ 0, \text{ if } j \to i \text{ based on prior knowledge, when } j < i. \end{cases} \tag{3.7}$$

Note that the utility of matrix C in the penalty corresponds to the objective of exploratory analysis, namely all available edges will be kept in the analysis. And $\Xi = \{\xi_{ij}\}$ is a $P \times P$ lower triangular matrix of adaptive weights with (i, j)-th element given by

$$\xi_{ij} = \begin{cases} \max(1, |\widetilde{\theta}_{ij}|^{-\gamma}), & \text{if } c_{ij} = 1 \text{ and } j < i; \\ 0, & \text{otherwise,} \end{cases} \tag{3.8}$$

where $\widetilde{\theta}_{ij}$ is obtained from the regular LASSO estimation obtained from (3.6) by setting $\xi_{ij} = 1$ if $c_{ij} = 1$ and $j < i$.

3.3.2 EM-Coordinate-Descent Algorithm

We propose a two-step iterative approach to estimate three unknown matrices (Θ, B, Ψ) in the SFEM. Given the current estimates $(B^{(t)}, \Psi^{(t)})$, $\Theta^{(t+1)}$ is updated by minimizing the penalized loss function (3.6) using the coordinated descent (CD) algorithm, and then $(B^{(t+1)}, \Psi^{(t+1)})$ are updated through the EM algorithm. Both EM algorithm and CD algorithm are presented below. Repeating these two-step procedure iteratively till convergence, we obtain estimates $(\hat{\Theta}, \hat{B}, \hat{\Psi})$ in the end.

3.3.2.1 EM algorithm

Similar to Section 2.4.1 in Chapter 2, the EM algorithm is used to estimate (B, Ψ) in the factor analysis model. We may implement the EM algorithm by treating the latent factors $z_n = (z_{n1}, \ldots, z_{nK})^T$, $n = 1, \ldots, N$ as "missing data" and Θ as a fixed

"known" constant matrix, where the M-step maximizes the joint normal log-likelihood of the full data $\{(y_n^* \triangleq y_n - \Theta y_n,\ z_n),\ n = 1, \ldots, N\}$. It is easy to derive the EM algorithm to update B and Ψ at the $(t+1)$-th iteration, respectively, given as follows.

In the **E-step**, we obtain the following moments of z_n, $n = 1, \ldots, N$,

$$\begin{aligned} E(z_n|y_n^*; B, \Psi) &= (B^T\Psi^{-1}B + I_K)^{-1}B^T\Psi^{-1}y_n^*, \\ Var(z_n|y_n^*; B, \Psi) &= I_K - B^T\Psi^{-1}B(B^T\Psi^{-1}B + I_K)^{-1}, \\ E(z_n z_n^T|y_n^*; B, \Psi) &= E(z_n|y_n^*; B, \Psi)E(z_n^T|y_n^*; B, \Psi) + Var(z_n|y_n^*; B, \Psi). \end{aligned} \tag{3.9}$$

In the **M-step**, B and Ψ are updated at the $(t+1)$-th iteration by, respectively,

$$\begin{aligned} B^{(t+1)} &= \Big\{\sum_{n=1}^{N} y_n^* E(z_n^T|y_n^*; B^{(t)}, \Psi^{(t)})\Big\}\Big\{\sum_{n=1}^{N}\Big[E(z_n z_n^T|y_n^*; B^{(t)}, \Psi^{(t)})\Big]\Big\}^{-1}, \\ \Psi^{(t+1)} &= \frac{1}{N}\sum_{n=1}^{N} E\Big[(y_n^* - Bz_n)(y_n^* - Bz_n)^T|y_n^*; B^{(t)}, \Psi^{(t)}\Big]. \end{aligned} \tag{3.10}$$

Specifically, if $\Psi = \sigma^2 I_Q$, we consider a simple re-parameterization by letting $\tilde{B} = \sigma^{-1}B$, and $\tilde{z}_n = \sigma z_n$. Clearly, Bz_n and $\tilde{B}\tilde{z}_n$ follow the same distribution. Thus, the EM algorithm updates $\tilde{B}$ and σ^2 at the $(t+1)$-th iteration by the following expressions:

$$\begin{aligned} \tilde{B}^{(t+1)} &= \Big\{\sum_{n=1}^{N} y_n^* E\Big(\tilde{z}_n^T|y_n^*; \tilde{B}^{(t)}, \sigma^{2(t)}\Big)\Big\}\Big\{\sum_{n=1}^{N} E\Big(\tilde{z}_n\tilde{z}_n^T|y_n^*; \tilde{B}^{(t)}, \sigma^{2(t+1)}\Big)\Big\}^{-1}, \\ \sigma^{2(t+1)} &= \frac{1}{NQ}\sum_{n=1}^{N} y^{*T}{}_n\Big\{I_Q - \tilde{B}^{(t)}\big(I_K + \tilde{B}^{T(t)}\tilde{B}^{(t)}\big)^{-1}\tilde{B}^{T(t)}\Big\}y_n^*. \end{aligned} \tag{3.11}$$

3.3.2.2 Coordinate descent algorithm

We implement an efficient algorithm to yield the optimal solution that minimizes the L_1-norm penalized loss function (3.6) under a fixed positive definite matrix $W = BB^T + \Psi$. Since minimizing (3.6) with respect to Θ is equivalent to a convex op-

timization problem, the objective function decreases over iterations, and the algorithmic convergence is warranted (*Tseng*, 2009). We first reformulate the optimization, so that the loss function (3.6) reduces to a regular LASSO regression problem with $\xi_{ij} = 1$ and $c_{ij} = 1$ in (3.6). Then, we apply the following active-shooting algorithm to find the sparse solution of Θ efficiently.

Given $Y^T_{P\times N} = (y^T_1, \ldots, y^T_N)$ and $\widetilde{Y} \triangleq YW^{-1/2}$ with $W = BB^T + \Psi$, it is easy to see that the quadratic loss function $\frac{1}{2N}\sum_{n=1}^{N}(y_n - \Theta y_n)^T(BB^T + \Psi)^{-1}(y_n - \Theta y_n)$ equals to $\frac{1}{2N} \parallel \mathcal{Y} - \mathcal{X}\beta \parallel^2$, where $\beta = (\theta_{21}, \ldots, \theta_{P1}, \ldots, \theta_{PP-1})^T$, $\mathcal{Y} = (\widetilde{Y}^T_2, \ldots, \widetilde{Y}^T_P)^T$, and $\mathcal{X} = (\mathcal{X}_{(2,1)}, \ldots, \mathcal{X}_{(P,P-1)})$ is an $N(P-1)$ by $\frac{P(P-1)}{2}$ matrix with $\mathcal{X}_{(i,\ j)} = (\underset{1^{\text{st}}\text{ block}}{0}, \ldots, \underset{(i-1)^{\text{th}}\text{ block}}{\widetilde{Y}^T_j}, \ldots, \underset{(P-1)^{\text{th}}\text{ block}}{0})^T$. Thus, the L_1-norm minimization in (3.6) is equivalent to the following optimization:

$$\min_{\Theta} \frac{1}{2N} \parallel \mathcal{Y} - \mathcal{X}\beta \parallel^2 + \lambda \sum_{h=1}^{\frac{P(P-1)}{2}} \xi_h |c_h \beta_h|, \tag{3.12}$$

with ξ_h being the h-th element of vector $\xi = (\xi_{21}, \ldots, \xi_{PP-1})^T$ and c_h being the h-th element of vector $c = (c_{21}, \ldots, c_{PP-1})^T$. The dimensions of $\mathcal{Y}$ and β are $N(P-1)$ and $P(P-1)/2$, respectively, which are higher than N and P. This could give rise to significant computational burden. Note that $\mathcal{X}$ is a block matrix with many zero blocks. Thus, taking such structural feathers into computation can help run the LASSO optimization algorithm more efficiently. To further boost the computational efficiency, we implement the active-shooting method (*Friedman et al.*, 2007, 2010a; *Peng et al.*, 2009) in the coordinate descent algorithm. It can be shown that the resulting computational complexity of solving (3.12) is $\min(O(NP^2), O(P^3))$, which is equivalent to performing P individual LASSO regressions in the neighborhood selection method (*Shojaie and Michailidis*, 2010).

Unlike the neighborhood selection method (*Shojaie and Michailidis*, 2010) which imposes sparsity on individual neighborhoods during optimization, the sparsity of

β is treated in a global fashion via the regularized objective function (3.12). Thus, our approach utilizes the data more efficiently, and appears more natural to deal with networks with hubs corresponding to, for example, master regulators. Indeed, detecting master regulators is of great interest in the reconstruction of gene regulatory networks. In addition, when certain *a priori* knowledge of directed edges is available, the proposed method (3.12) has the flexibility of incorporating such prior knowledge. For example, we can determine whether to penalize a pair of nodes by including the corresponding entry in the weight term c or not. Also with the utilization of the term ξ, we can assign different adaptive weights to different pairs of nodes according to their importance.

The active-shooting algorithm proceeds as follows: at each updating step, we first define an "active" set of currently nonzero coefficients and update the coefficients within the active set until convergence is achieved before moving on to update other parameters. This is feasible because the active set usually remains small under the sparse model assumption. Defining a current active set $H = \{h : c_h\beta_h \neq 0\}$, we update $\beta_{h_0 \in H}$ by (3.13) with all other $\beta_{h \neq h_0}$ fixed until convergence is achieved in H.

$$\hat{\beta}_{h_0} = \begin{cases} (\mathcal{Y} - \sum_{h \neq h_0} \beta_h \mathcal{X}_h)^T \mathcal{X}_{h_0} / \|\mathcal{X}_{h_0}\|_2^2, \text{ if } c_{h_0} = 0, \\ S\Big((\mathcal{Y} - \sum_{h \neq h_0} \beta_h \mathcal{X}_h)^T \mathcal{X}_{h_0},\ N\lambda\xi_{h_0}\Big) / \|\mathcal{X}_{h_0}\|_2^2, \text{ if } c_{h_0} = 1, \end{cases} \tag{3.13}$$

where $S(a, b) = sgn(a)(|a| - b)_+$ is the soft-thresholding operator.

Finally, a combination of the EM algorithm and the CD algorithm, termed as the EM-CD algorithm, allows us to iteratively update Θ, B and Ψ. The detail is provided in the following Algorithm 4.

Algorithm 4 EM-CD algorithm

Step 1. Initialization of $B^{(0)}$, $\Psi^{(0)}$, and $\beta^{(0)}$ with some suitable values
Step 2. Given $B^{(t)}$, $\Psi^{(t)}$, and $\beta^{(t)}$, for iteration t+1, $\beta^{(t+1)}$ is updated by the active-shooting CD algorithm
Step 3. Given $\beta^{(t+1)}$, ($B^{(t+1)}$, $\Psi^{(t+1)}$) are iteratively updated based on the EM algorithm till convergence
Step 4. Repeat the above two steps till convergence.

3.3.3 Tuning parameter selection

The choice of the number of latent factors K and the tuning parameter λ are of great importance in the proposed method. We first consider the selection of the number of latent factors K, and then discuss the selection of tuning parameter λ. The number of latent factors in the proposed SFEM can affect the resulting sparsity in the weighted adjacency matrix Θ and have to be tuned properly. Because the SFEM is a generalized type of factor analysis model, methods that have been developed in the literature for selecting the number of factors may be applied to the SFEM. *Bai and Ng* (2002) and *Onatski* (2010) proposed statistics to determine the number of static factors in certain approximate factor analysis models. *Onatski* (2009) developed tests for the number of factors using the empirical distribution of eigenvalues of the sample covariance matrix. *Hirose and Konishi* (2012) and *Caner and Han* (2013) applied the shrinkage estimation to select relevant factors.

In this chapter, we propose to use an "eigenvalue ratio (ER)" criterion to select the number of latent factors K, due mainly to its simplicity and computational ease (*Ahn and Horenstein*, 2013). Note that the generalized factor model $y = (I - \Theta)^{-1}Bz + (I - \Theta)^{-1}e = \Gamma z + \delta$ given in (3.5), where $\Gamma = (I - \Theta)^{-1}B$. Hence, we can covert the selection of the number of latent factors K in B to select K in Γ based on the generalized factor model. Following *Ahn and Horenstein* (2013), for a sample covariance matrix $YY^T/(NP)$, denote its k^{th} largest eigenvalues by η_k, $k = 1, \ldots, \min(N, P)$. The corresponding eigenvalue ratio is

given by $\mathrm{ER}(k) = \eta_k/\eta_{k+1}$. Then we propose an "eigenvalue ratio" criterion as $K_{\mathrm{ER}} = \arg\max_{K_{min}\le k\le K_{max}} \mathrm{ER}(k)$, where K_{min} and K_{max} may be prespecified by the scree plot; for example $K_{min} = 1$ or 2, $K_{max} = min(N, P)/2$.

In respect to the selection of tuning parameter λ, we adopt the M-fold cross-validation method. Since the true model is believed to be sparse, we utilize the ordinary least squares (OLS) estimates instead of the shrunken estimates to calculate the cross-validation score. This is because, when there are many potential poor predictors, the cross-validation score based on the shrunken estimates often leads to severe false positive rates (*Peng et al.*, 2010; *Efron et al.*, 2004). In contrast, using the OLS estimates seems to make a reasonable remedy for such a problem, which is also observed in our simulation studies. It is worth pointing out that Bayesian information criterion (BIC), another popular tuning selection method, is not considered here, mainly because estimating the degrees of freedom required in the BIC is difficult under a nonorthogonal design.

3.4 Numerical Results

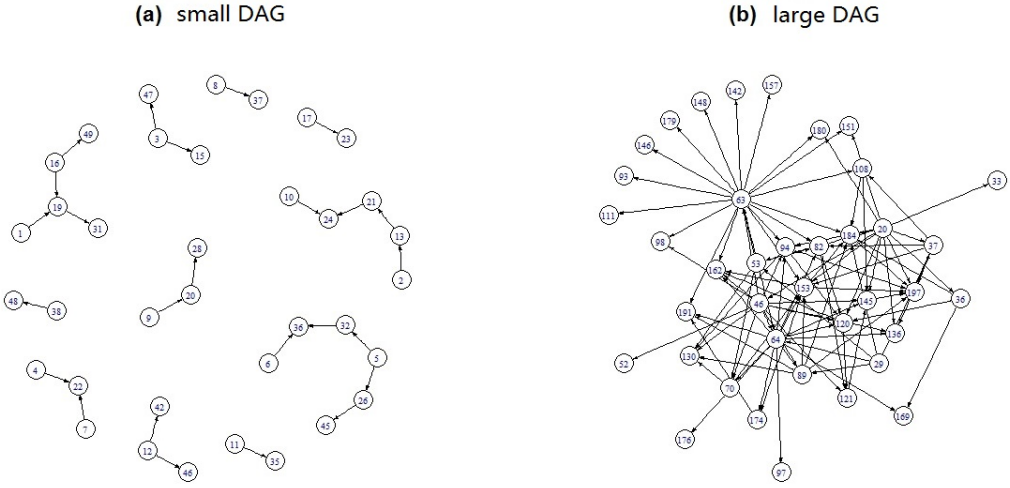
To examine the performance of the proposed SFEM for the exploratory analysis, we conduct two simulation experiments.

The two simulated DAGs are displayed in Figure 3.3 with details given as follows. In simulation experiment I, we begin with a small DAG, which is randomly generated with $P = 50$ nodes and $M = 25$ edges by the R-package *pcalg* (*Kalisch and Buhlmann*, 2007). To control for the sparsity of the graph, we further set the maximum number of parents for any given node is 2 and the depth of DAG is 3. Then we randomly generate DAG until exact $M = 25$ edges are achieved. Figure 3.3(a) displays one simulated DAG.

In simulation experiment II, we investigate a more complex DAG consisting of 19 master regulators (i.e. parental nodes). Among them, 4 are strong master regulators,

each influencing 14 to 18 nodes, 7 are weak master regulators, each influencing 3 to 7 nodes, and the rest 8 parental nodes link to only 1 or 2 nodes. Such DAG topology is generated by randomly selecting 19 master parental nodes, and within each parental node, children nodes are further randomly selected. As a result, we create a DAG with $M = 100$ edges. In this second experiment, we begin with the SFEM with $P = 200$ nodes and different number of latent factors $K = 1, 5, 10$. For the case of $K = 5$, we vary the number of nodes as $P = 50, 100, 200$. Clearly, with a fixed number of edges $M = 100$, the sparsity of DAG increases with an increased P. Figure 3.3(b) shows a simulated DAG.

Figure 3.3: Topology of two simulated DAGs. (a) A small DAG with 50 nodes and 25 edges. (b) A large DAG with 200 nodes and 100 edges.



For each of the two DAGs we generate $N = 25,\ 100$ observations, respectively, from the structural factor equation model (3.4). For the networks of these directed edges in Figure 3.3, we generate adjacency weights $\theta_{ij} \overset{i.i.d.}{\sim} U([-3,-1]\bigcup[1,3])$ in Simulation I, and set constant $\theta_{ij} = 0.5$ in Simulation II. In each case, we simulate latent factors $z_{nk} \overset{i.i.d.}{\sim} N(0,1)$, loadings $B_{ik} \overset{i.i.d.}{\sim} U([-b,-a]\bigcup[a,b])$ and noise $e_{nj} \overset{i.i.d.}{\sim} N(0,\sigma^2)$, where a, b and σ^2 are chosen to satisfy a prespecified signal-to-noise ratio: $SNR = \sqrt{tr(\Sigma)/tr(\Sigma_\delta)}$. The tuning parameter is determined by the 5-fold cross validation method, and 50 replicates are carried out to draw summary statistics.

The performance of the exploratory analysis via the SFEM is mainly compared under three scenarios: (i) ignoring latent factors ($K = 0$, i.e. *Shojaie and Michailidis* (2010)); (ii) the number of latent factors K is over/under-specified; and (iii) the number of latent factors K is unknown but selected by the eigenvalue-ratio (ER) method, i.e. $K = \hat{K}_{ER}$. Note that $K_{\text{ER}} = \max_{K_{min} \le k \le K_{max}} \eta_k/\eta_{k+1}$, where η_k is the k^{th} largest eigenvalue of $\sum_{n=1}^{N} y_n y_n^T/(NP)$. K_{min} and K_{max} are chosen as $K_{min} = 1$, $K_{max} = min(N, P)/2$ in the simulation studies.

For each simulated dataset, we generate the solution path for Θ using a geometric sequence of tuning parameter λ's, starting from the largest value λ_{max} for which $\widehat{\Theta}_{\lambda_{\text{max}}} = \mathbf{0}$ and decreasing to the smallest value $\lambda_{\text{min}} = 10^{-4}$. Note that the number of totally detected edges increases as tuning parameter λ decreases. We then evaluate the performance of the SFEM under different latent factors nested within a series of the tuning parameter values. In addition, we further compare the performance of the proposed SFEM with the existing PC-algorithm (*Kalisch and Buhlmann*, 2007), where the significance levels of the PC-algorithm are given by a geometric sequence of values ranging in $[10^{-10}, \ldots, 0.95]$. Note that the PC-algorithm does not require the node ordering as an input. So to be fair, in this comparison we simply ignore the direction of an edge and a correct discovery includes either an directed edge or undirected edge.

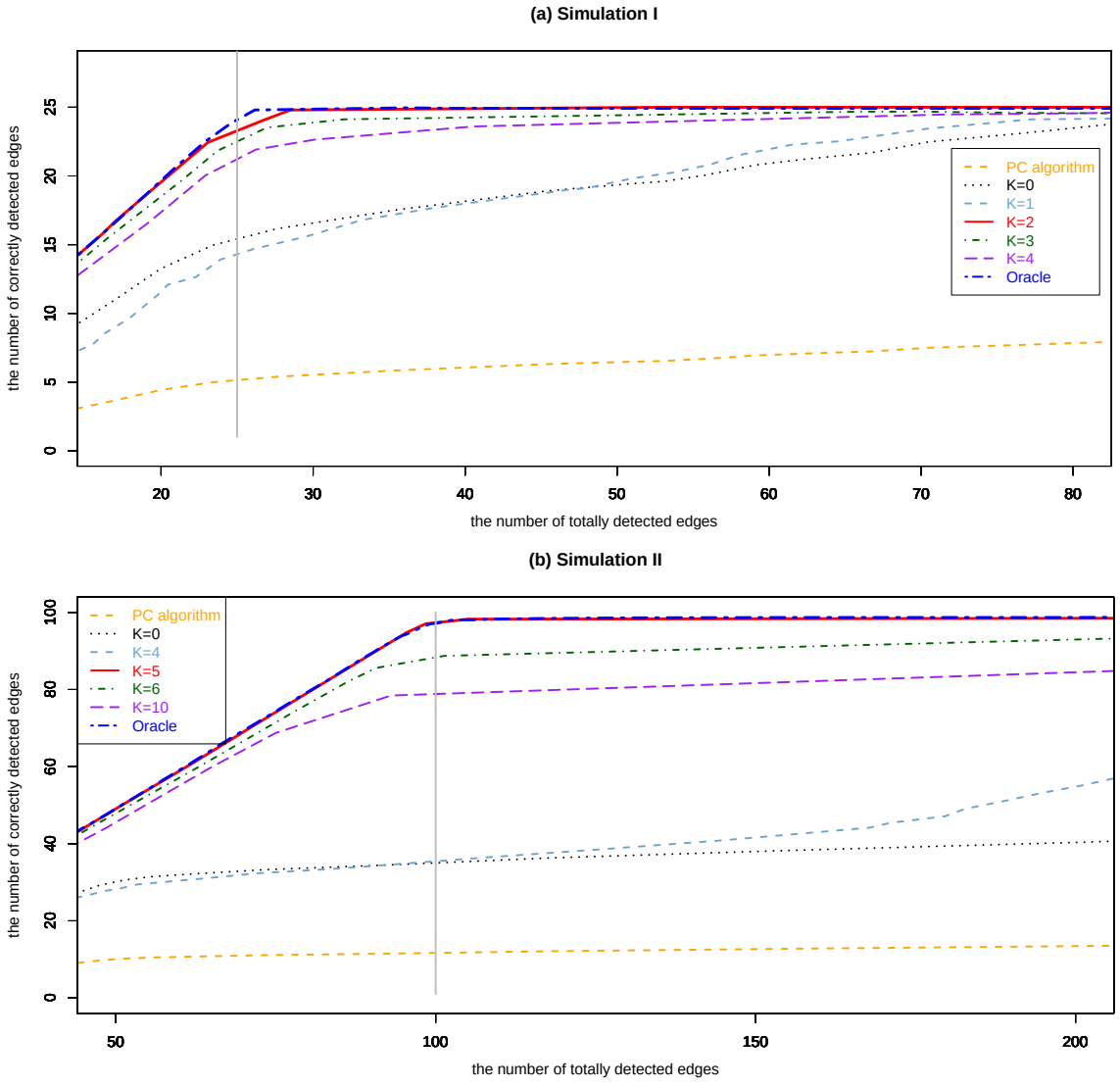
Figure 3.4 shows the plot of the number of correctly detected edges versus the number of totally detected edges across different numbers of latent factors averaged over 50 replicates. Here "oracle" represents the case where the SFEM uses the true covariance matrix $W = BB^T + \Psi$ without estimation of B and Ψ. We observe that the case of $K = K_{\text{true}}$ with estimated B and Ψ outperforms the all other cases with a misspecified K, and its performance is close to that of the "oracle" case. This means that the EM algorithm works well to estimate W matrix. The PC-algorithm performs the worst, and is even worse than the SFEM with $K = 0$. This is probably because

the PC-algorithm does not utilize any *a priori* knowledge of node ordering and is operated under the setting of completed partially directed acyclic graph (CPDAG). As a consequence, it may include many false undirected edges induced by the shared latent factors. Specifically, in Figure 3.4(b) when we detect 100 edges, the sparse SFEM with $K = K_{\text{true}}$ can detect more than 95% of the true edges correctly with an average standard deviation of 1.45 edges, whereas the PC-algorithm can only detect about 10% of the true edges successfully.

The quality of our method is further measured by the average number of true positive(TP), false positive(FP), false negative(FN) edges, and sensitivity(Sen) and Matthews correlation coefficient score (MCC). Table 3.1 summarizes the average performance of the SFEM with $K = K_{ER}$ for different number of P in Simulation experiment II with $K_{\text{true}} = 5$. For example, when $P = 200$ and $\theta_{ij} = 0.5$, on average the estimated graph is able to identify 104.04 directed edges, of which 98.12 edges are the true edges, and in Θ the other 5.92 edges are false. Note that when $P = 200$, the number of parameters to be estimated is around $20,000$, which is much larger than the sample size $N = 100$. In this high-dimensional setting with a substantial influence of latent factors ($K = 5$), results in Table 3.1 suggest that our regularization method can estimate the DAG structure with reasonable accuracy even with the limited sample size $N = 100$. Of course, when P is relatively small (i.e. $P = 50$ or 100), the SFEM method shows better results than those with $P = 200$.

Table 3.2 lists the results of Simulation experiments I and II with different numbers of latent factors. Table 3.2 suggests that the proposed ER criterion works quite well in selecting the number of latent factors, except for the case of Simulation II with SNR=2:1. This is because in this setting SNR is relatively small, and it is interesting to notice that, although $K = 2$ is selected, which is near the true $K = 1$, the resulting performance ($K_{ER} = 2$) appears much better than that with an underspecified $K = 0$ (e.g. MCC=0.85 versus 0.29). As shown in the Table 3.2, ignoring (or

Figure 3.4: Simulation results two DAG networks, where x-axis is the number of totally detected edges, and y-axis is the number of correctly identified edges. The vertical grey line corresponds to the number of true edges. (a) Simulation I: $P = 50, N = 25, K = 2, M = 25$, and $K_{ER} = 2$. (b) Simulation II: $P = 200, N = 100, K = 5, M = 100$, and $K_{ER} = 5$.



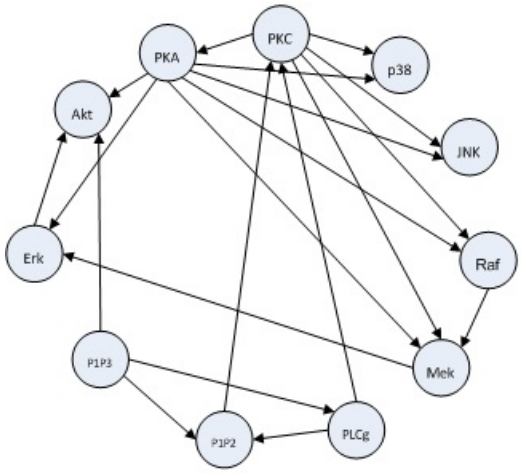
under-specifying) unobserved factors in the SFEM results abundant nonzero entries in Θ that produce excessive false edges. In contrast, if the number of factors is over-identified, the resulting SFEM would produce a scant Θ matrix that leads to many false negative discoveries. To sum up, the proposed SFEM$_{ER}$ method shows a satisfactory performance with the highest sensitivity and MCC as well as the lowest total false rate.

3.5 Analysis of cell signaling data

This section demonstrates an application of the proposed SFEM method to analyze multivariate flow cytometry data available in *Sachs et al.* (2005), which has been previously analyzed by *Shojaie and Michailidis* (2010); *Fu and Zhou* (2013); *Friedman et al.* (2008); *Aragam and Zhou* (2014), among others. This dataset includes 11 phosphorylated proteins from $N = 7466$ cells. The consensus network, constructed by experimental annotations, has 20 edges, which is displayed in Figure 3.5 and is used as the benchmark to assess the accuracy of an estimated network structure. A direction from node i to node j is interpreted as a causal influence from protein i to protein j. Following *Shojaie and Michailidis* (2010), the node ordering in the DAG is treated as *a priori* feature among 11 proteins.

Based on the scree plot in Figure 3.6 and the eigenvalue-ratio (ER) method, we obtain $K_{ER} = 4$. We explore the SFEM under different numbers of latent factors $K = 0, 2, 4, 6$, where $K = 0$ corresponding to the analysis given by *Shojaie and Michailidis* (2010) and $K = 4$ is the estimated number of latent factors. Figure 3.7 shows the plot of the number of correctly detected edges versus the number of totally detected edges across different number of latent factors. Compared with $K = 0, 2, 4, 6$, we find that the SFEM with the estimated $K_{\text{ER}} = 4$ performs slightly better than the other cases. To compare the SFEM with $K = 0$ with the SFEM with $K = 4$ when both methods detected 25 edges, out of them, 10 edges detected by the latter are

Figure 3.5: The consensus signaling network of 11 proteins.



in the consensus network, while 7 edges detected by the former are in the consensus network. For these two models, after selecting the optimal tuning parameters from the 5-fold cross-validation method, compared with the SFEM with $K = 0$, we find that the SFEM with $K = 4$ shows a different DAG in Figure 3.8, the difference occurs mainly in the domain of false discoveries. For example, the SFEM with $K_{ER} = 4$, which adjusts for 4 shared latent factors, performs better than the SFEM with $K = 0$ in terms of low FP; among the total of 25 directed edges, 15 edges from $K_{ER} = 4$ are false positive in comparison to 18 false positive edges from $K = 0$. Hence, the SFEM with $K_{ER} = 4$ appears more reliable in the data analysis, which gives fewer false positives and more true positive signals in comparison to the SFEM with $K = 0$.

Nevertheless, several known edges are not detected by both SFEMs with $K = 0$ and $K_{ER} = 4$. One possible reason is that the proposed SFEM is a linear Bayesian network, which may not be able to detect nonlinear causal relationships.

Figure 3.6: The scree plot of eigenvalues.

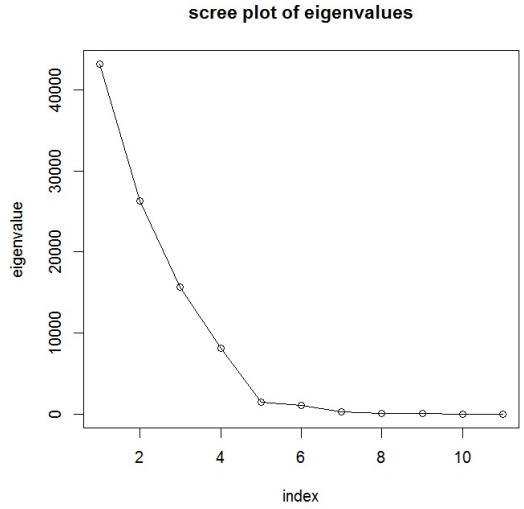


Figure 3.7: The plot of the correct discovery, where x-axis is the number of totally detected edges, and y-axis is the number of correctly identified edges.

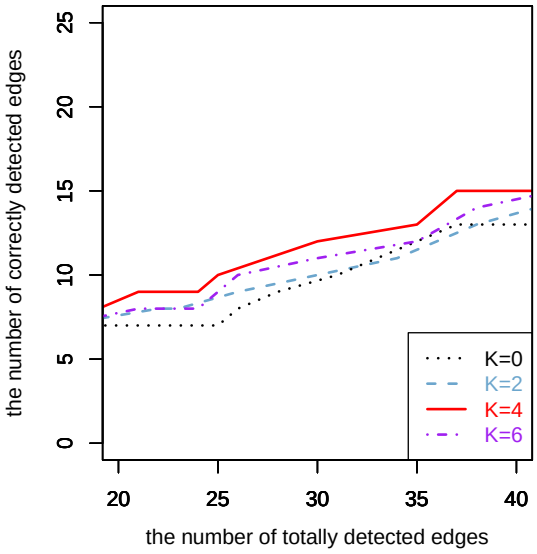
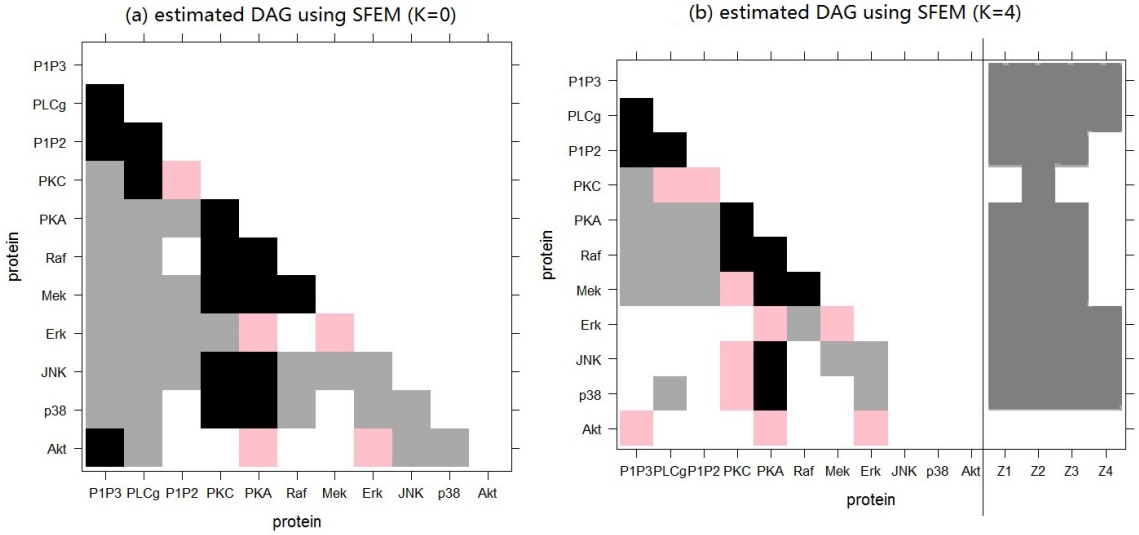


Figure 3.8: Causal interactions among 11 proteins of the signaling pathway: black represents TP, pink represents FN, and grey represents FP. In the right panel, $Z_1, \ldots, Z_4$ represents 4 common latent factors.



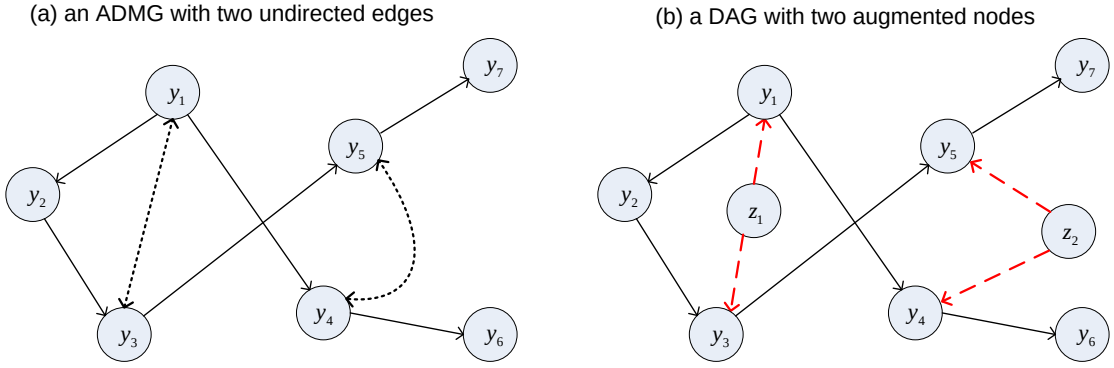
3.6 Confirmatory analysis of ADMG

3.6.1 Formulation

As mentioned above, a confirmatory analysis concerns the situation where there is the knowledge about the existence of undirected edges together with directed edges. In this case of ADMG, the proposed sparse SFEM may be used to reconstruct the subgraph of directed edges, accounting for undirected edges. The key idea behind the use of SFEM to deal with edges of mixed types stems from the augmentation approach that converts each undirected edge into two directed edges via an augmented variable as their parental node. The resulting graphical model is known as the semi-Markovian causal model (SMCM), in which $y_i \leftrightarrow y_j$ is replaced by a new path $y_i \leftarrow z^{(i,j)} \rightarrow y_j$, where $z^{(i,j)}$ is the augmented variable, see for example *Pearl* (2000); *Richardson and Spirtes* (2002); *Kalisch and Bhlmann* (2013). In this way, we transform an ADMG into a DAG by using augmented parental nodes to enlarge the vertex set. Then the proposed SFEM is ready to be applied to model ADMG with the augmented nodes

being treated as latent factors. Figure 3.9 shows an example, in which two augmented nodes z_1 and z_2 are introduced in Figure 3.9(b) to reformulate two undirected edges $y_1 \leftrightarrow y_3$ and $y_4 \leftrightarrow y_5$.

Figure 3.9: Graphical representation of SFEM for confirmatory analysis.



By enlarging the vertex set V from $V = (y_1, \ldots, y_7)$ to $V = (y_1, \ldots, y_7, z_1, z_2)$, we can jointly model the above ADMG in Figure 3.9 (a) by the structural equation model with two augmented nodes:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{11} & 0 \\ \theta_{21} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{32} & 0 & 0 & 0 & 0 & 0 & b_{31} & 0 \\ \theta_{41} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{42} \\ 0 & 0 & \theta_{53} & 0 & 0 & 0 & 0 & 0 & b_{52} \\ 0 & 0 & 0 & \theta_{64} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{75} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \\ \delta_7 \\ \delta_8 \\ \delta_9 \end{bmatrix}. \tag{3.14}$$

Note that the weighted adjacency matrix in (3.14), denoted by $\widetilde{\Theta}$, is a block matrix

given as follows:

$$\widetilde{\Theta} = \begin{bmatrix} \Theta_{7\times 7} & B_{7\times 2} \\ 0_{2\times 7} & 0_{2\times 2} \end{bmatrix}.$$

By assuming that $\delta_y = (\delta_1, \ldots, \delta_7)^T \sim \text{MVN}_7(0, \Psi)$ and $z = (z_1, z_2)^T = (\delta_8, \delta_9)^T \sim \text{MVN}_2(0, I)$, it is easy to show that the covariance of $y = (y_1, \ldots, y_7)^T$ is $\Sigma = (I - \Theta)^{-T}(BB^T + \Psi)(I - \Theta)^{-1}$, which is the same as the covariance matrix given in the SFEM (3.4). Thus, by treating augmented variables as latent factors, we can apply the proposed SFEM to model ADMG in Figure 3.9 (a) by

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{21} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{32} & 0 & 0 & 0 & 0 & 0 \\ \theta_{41} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{53} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_{64} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{75} & 0 & 0 \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} + \begin{bmatrix} b_{11} & 0 \\ 0 & 0 \\ b_{31} & 0 \\ 0 & b_{42} \\ 0 & b_{52} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \times \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \\ \delta_7 \end{bmatrix}, \tag{3.15}$$

where the position of nonzero elements in the factor loading matrix B is predetermined by induced augmented variables, and we call these nonzero b_{ij}'s as unconstrained parameters, whereas these zero b_{ij}'s as constrained parameters. Furthermore, we call the above SFEM (3.15) as the SFEM for confirmatory analysis (SFEM-CA). Note that the SFEM-CA has the same expression as the SFEM given in (3.4). However, the SFEM-CA considers a structured factor loading matrix B via *a priori* knowledge of undirected edges, which is different from the unstructured B matrix included in the SFEM for exploratory analysis.

Recall that the SFEM is a generalized factor model and can be rewritten as $y = (I - \Theta)^{-1}Bz + (I - \Theta)^{-1}e = \Gamma z + \delta$. For the SFEM-CA, we impose extra assumptions to identify $\Sigma_\delta = (I - \Theta)^{-1}\Psi(I - \Theta)^{-T}$ and Γ separately. Note that the identification condition for Σ_δ in SFEM-CA is the same as the identifiability condition

(B) given in Section 3.2.4. Then we impose another assumption as follows:

- Identifiability condition (C) for Γ: each factor has at least three children (*Grzebyk et al.*, 2004)

Note that the identifiability condition (C) is a necessary condition for identification of a generalized factor model (*Grzebyk et al.*, 2004), and this condition is widely used for its ease of verification (*Snchez et al.*, 2005). It is easy to show that the SFEM displayed in Figure 3.9(b) also satisfies this condition, since each column of Γ has at least three nonzero entries. One can refer to *Grzebyk et al.* (2004) and *Drton et al.* (2011) for more theoretical discussions about the sufficient and necessary conditions for identifiability of a generalized factor model and conditions for global identifiability of ADMGs using linear SEM, respectively.

Similar to Section 3.3.2, the EM algorithm is used to estimate (B, Ψ) in the SFEM-CA. The E-step in (3.9) is the same for the SFEM-CA. The M-step that maximizes the normal joint log-likelihood of the full data $\{(y_n^* \triangleq y_n - \Theta y_n,\ z_n),\ n = 1, \ldots, N\}$ is different. In the M-step for the SFEM-CA, $B = (b_1, \ldots, b_P)^T$ and $\Psi = \text{diag}(\Psi_1, \ldots, \Psi_P)$ are updated at one variable y_p at one time $p = 1, \ldots, P$. Consider the p-th variable y_p with factor loading vector $b_p = (b_{p1}, \ldots, b_{pK})^T$ on K latent factors. Rearrange the elements of b_p as $U_p b_p = [b_p^{1^T}, b_p^{0^T}]^T$ by a $K \times K$ rotation matrix U_p, where b_p^1 is a vector that contains K_p unconstrained loadings to be estimated, and b_p^0 is a vector that consists of $K - K_p$ constrained loadings equal to zero. Accordingly, similar partitions are applied to a $K \times K$ matrix $Q \triangleq \sum_{n=1}^N E(z_n z_n^T | y_n^*; B, \Psi)$ and the p-th row of a $P \times K$ matrix $\eta \triangleq \sum_{n=1}^N y_n^* E(z_n^T | y_n^*; B^{(t)}, \Psi^{(t)})$ in the E-step, so the resulting $K_p \times K_p$ block Q_{1p} and $1 \times K_p$ block η_{1p} correspond to the subvector of unconstrained loadings for variable y_p. The M-step updates the subvector b_p^1 and

uniqueness Ψ_p as follows:

$$\begin{aligned} b_p^1 &= Q_{1p}^{-1}\eta_{1p}^T, \\ \Psi_p &= S_p - \eta_{1p}Q_{1p}^{-1}\eta_{1p}^T \end{aligned} \tag{3.16}$$

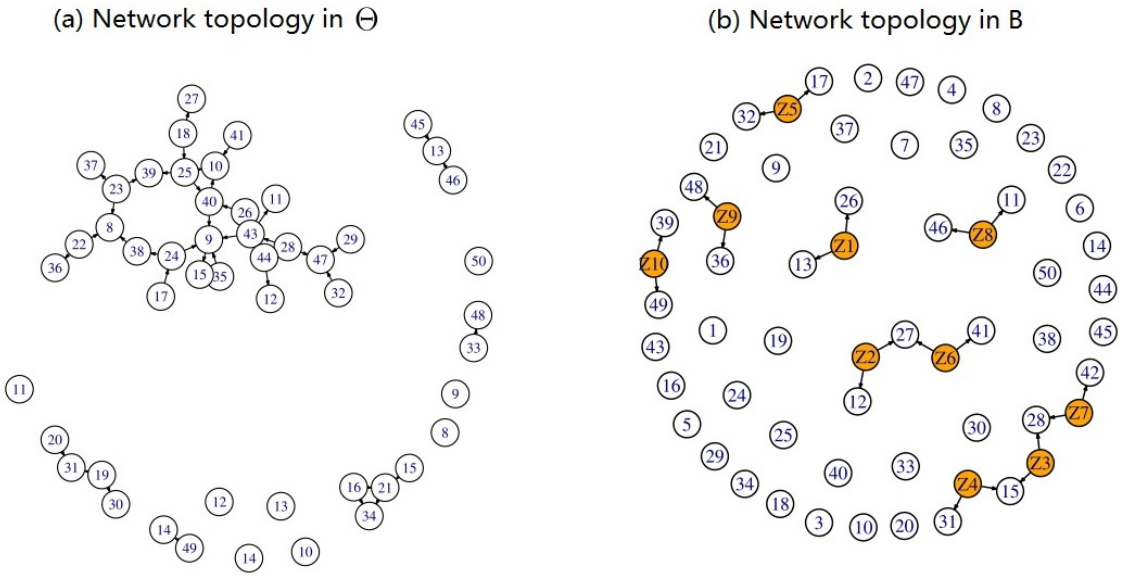
where S_p is the p-th diagonal element of $S = \sum_{n=1}^{N} y_n^* y_n^{*T}$. Furthermore, we can implement the EM-CD Algorithm 4 in a similar way to estimate Θ by optimizing the objective function (3.6) or equivalently(3.12).

3.6.2 Some numerical results

We assess the performance of the proposed SFEM-CA by a simulation experiment. We first randomly generate a DAG with $P = 50$ nodes and $M = 40$ edges by the R-package *pcalg* (Kalisch et al., 2007). To control for the sparsity, we set the maximum number of parents for any given node as 4. Then, we add 10 randomly generated undirected edges. To guarantee the parameter identifiability in the SFEM-CA, we restrict each column of the generated $\Gamma = (I - \Theta)^{-1}B$ to have at least three nonzero entries. One of the simulated ADMG is displayed in Figure 3.6.2. Observations are generated according to the SFEM-CA. We first generate $\theta_{ij} \overset{i.i.d.}{\sim} U([-3,-1]\bigcup[1,3])$ based on the topology among 50 observed variables shown in Figure 3.10 (a), 10 augmented variables $z_{nk} \overset{i.i.d.}{\sim} N(0,1)$, and noise $e_{nj} \overset{i.i.d.}{\sim} N(0,1)$. Given the relationships between 50 observed variables and 10 augmented variables shown in Figure 3.10 (b), we further generate factor loadings with a constant 0.5, where signs are randomly generated with probability 0.5. 50 replicates are carried out to draw summary statistics.

Similar to Section 3.4, for each simulated dataset, we generate the solution path along a geometric sequence of tuning parameter λ, starting from the largest value $\lambda_{\max}$ for which $\widehat{\Theta}_{\lambda_{\max}} = 0$ and decreasing to the smallest value $\lambda_{\min} = 10^{-3}$. To compare the proposed SFEM-CA with the classic PC-algorithm, a geometric sequence of the significance levels ranging in $[10^{-5}, \ldots, 0.8]$ is considered for the PC-algorithm. The

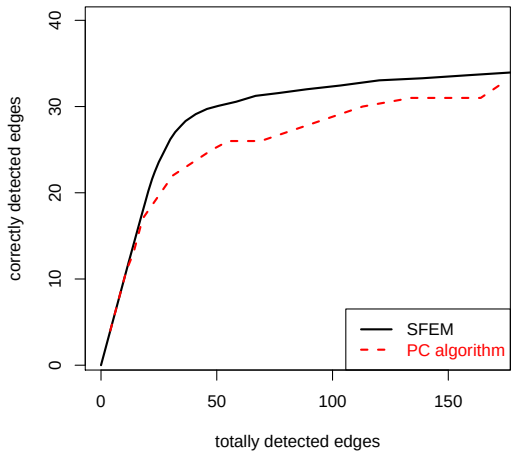
Figure 3.10: The network of an ADMG with 50 nodes and 50 edges. Among 50 edges, 40 are directed edges showed in (a) and 10 are undirected edges induced by 10 augmented variables $Z_1, \ldots, Z_{10}$ showed in (b).



performances of the proposed SFEM-CA and the PC-algorithm on the ADMG are shown in Figure 3.11. The plot shows the number of correctly detected edges versus the number of totally detected edges, averaged across 50 replicates. It is evident that the proposed SFEM-CA outperforms consistently over the PC-algorithm. For the example of both SFEM-CA and PC algorithm having detected 40 edges on average, approximately 29 edges given by the SFEM-CA are the true signals, whereas only 23 edges given by the PC-algorithm are correct. As mentioned above, since the PC-algorithm does not utilize any *a priori* knowledge of node ordering, it seems to perform better in the setting of ADMG than in a DAG whose edges may be contaminated by unmeasured factors considered in the exploratory analysis. The observed discrepancies between the SFEM-CA and the PC-algorithm in Figure 3.11 is marginal and largely attributed to the use of node ordering information in the SFEM-CA.

Based on the optimally chosen tuning parameter from the 5-fold cross-validation

Figure 3.11: The plot of the correct discovery in the ADMG over 50 replicates.



method, we compare this optimal SFEM-CA with the best case (in terms of the highest MCC) obtained from the PC algorithm with the significance level $\alpha = 0.05$. The corresponding results are summarized in Table 3.4. From this table, it is easy to see that the PC-algorithm tends to estimate a larger DAG with more false positives, whereas the SFEM-CA favors a smaller DAG with fewer false positives and the precision rate (PPV $= \frac{TP}{TP+FP}$) is around 92%.

3.7 Discussion

We have proposed a class of SFEMs for both exploratory analysis and confirmatory analysis with the availability of node ordering among the variables. The proposed new methodology is based on a combination of the structural equation model and the factor analysis model. The proposed SFEM may be regarded as a generalized factor analysis model that can separate directed and undirected edges by modeling the concentration matrix. Our presentation of the proposed model has been primarily based on the exploratory contexts.

When there are no latent factors included, the proposed SFEM reduces to the classical SEM. Thus, the reconstruction of DAGs based on our the proposed L_1 norm regularization method is equivalent to the L_1 norm penalized likelihood method

proposed by *Shojaie and Michailidis* (2010). From their simulation studies, Shojaie and Michailidis (2010) have shown that their method (i.e. the SFEM with $K = 0$) is not sensitive to random permutations of the order of variables in high dimensional sparse settings. However, their results depend on the fact that the randomly generated DAGs have a moderate size of *v-structures*(i.e., $i \to k \leftarrow j$). If the number of *v-structures* increases, the performance of any method utilizing a given node ordering would rapidly deteriorate (*Altomare et al.*, 2013). Hence, learning the order of the variables is of great importance. Recently, *Fu and Zhou* (2013), and *Aragam and Zhou* (2014) applied L_1-norm penalty and MCP penalty, respectively, to estimate DAGs from penalized likelihood under an unknown order of variables. However, their objective functions are non-convex, which might cause multiple local solutions in the optimization. Hence, it is interesting to explore how the direction of causality among network nodes may be possibly estimated under the SFEM in the future work. Also, when the order of variables is known *a priori*, the reconstruction of a causal network from time-course observations based on Granger causality (*Granger*, 1969) is a potentially promising area of research to extend the proposed SFEM method. In addition, in the real data analysis, many causal relations in gene regulatory networks are possibly nonlinear, which may not be detectable using the linear SFEM proposed in this chapter. Thus, learning nonlinearity casuality is another interesting extension of this research topic.

Table 3.1: Performance comparison under different number of nodes.

P	Total(TP+FP)	TP	FP	FN	Sen	MCC	K_{ER}(%)
50	99.07	97.12	1.95	1.94	0.98	0.99	5 (100%)
100	100.69	97.56	3.13	2.17	0.97	0.97	5 (100%)
200	104.04	98.12	5.92	1.88	0.98	0.96	5 (100%)

Table 3.2: Results of Simulation I and II: Impact of different number of latent factors K and different SNR levels on DAG estimation.

SNR	K_{true}	Method	Total(TP+FP)	TP	FP	FN	Sen	MCC	K_{ER}(%)
Simulation I									
3:1	2	SFRM_{ER}	29.44	24.76	4.68	0.24	0.99	0.92	2 (100%)
		$\text{SFEM}_{K=0}$	118.84	24.96	93.88	0.04	0.99	0.44	
		$\text{SFEM}_{K=5}$	25.88	20.24	5.64	4.76	0.81	0.80	
Simulation II									
4:1	5	SFEM_{ER}	104.04	98.12	5.92	1.88	0.98	0.96	5 (100%)
		$\text{SFEM}_{K=0}$	1530.92	97.96	1432.96	2.04	0.98	0.24	
		$\text{SFEM}_{K=7}$	72.29	63.86	8.43	36.14	0.64	0.70	
2:1	1	SFEM_{ER}	88.2	79.84	8.34	20.16	0.80	0.85	2(100%)
		$\text{SFEM}_{K=1}$	104.44	97.44	7.00	2.56	0.97	0.95	
		$\text{SFEM}_{K=0}$	1015.32	93.64	921.68	6.36	0.94	0.29	
6:1	10	SFEM_{ER}	93.76	91.52	2.24	8.48	0.92	0.94	10 (100%)
		$\text{SFEM}_{K=0}$	3686.72	97.64	3589.08	2.36	0.98	0.14	
		$\text{SFEM}_{K=15}$	53.28	49.64	3.64	50.36	0.50	0.67	

Table 3.3: Comparison between SFEM with K=0 and 4 ($K_{\text{ER}} = 4$) under the selected optimal tuning parameter.

Method	Total	TP	FP	FN	Method	Total	TP	FP	FN
K=0	42	15	27	5	K=4	25	10	15	10

Table 3.4: Comparison between the optimal SFEM and the best case of the PC-algorithm, where PPV denotes discovery precision rate (%).

Method	Total(TP+FP)	TP	PPV	FP	FN	Sen	Spec	MCC
SFEM	26.28	24.08	92	2.20	15.92	0.60	0.998	0.74
PC-algorithm	31.12	21.92	70	11.20	18.01	0.55	0.99	0.61

CHAPTER IV

Regression analysis of networked data

This chapter concerns the development of a new regression analysis methodology to assess relationships between multi-dimensional response variables and covariates that are correlated through networks. To address analytic challenges pertaining to the integration of network topology into the regression analysis, we propose a method of hybrid quadratic inference functions (HQIF) that utilizes both prior and data-driven correlations among network nodes into statistical estimation and inference. Moreover, a Godambe information based tuning strategy is proposed to allocate weights between the prior and data-driven pieces of network knowledge, so that the resulting estimation achieves desirable efficiency. The proposed method is conceptually simple and computationally fast, and more importantly has appealing large-sample properties in both estimation and inference. This new methodology is evaluated through simulation studies and illustrated by a motivating example of neuroimaging data about an association study of iron deficiency on infant's auditory recognition memory.

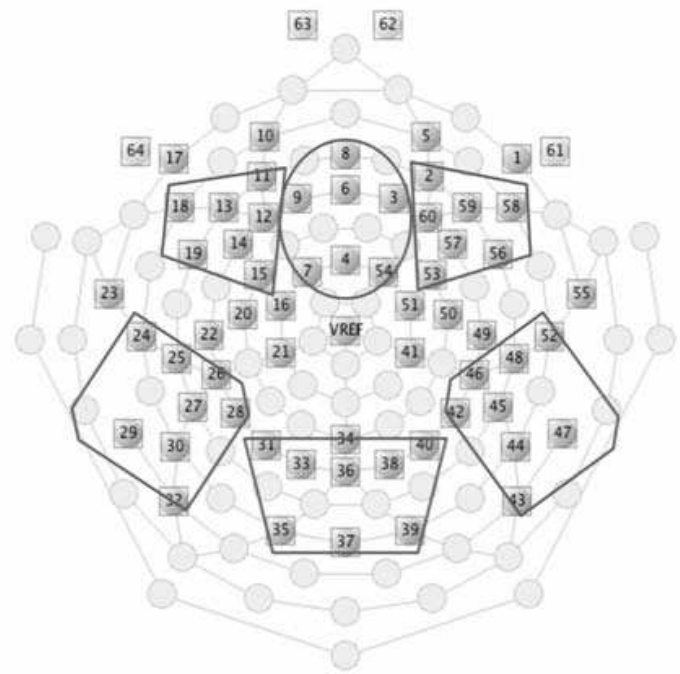
4.1 Introduction

Data collected from networks are pervasive in practice. A network refers to a set of nodes or vertices joined in pairs by edges (*Newman*, 2010). An important feature of a network is that between-node distance may not be defined precisely in a numeric

metric, and because of this it differs from the space-time system. The focus of this chapter is to develop a new methodology for regression analysis of multi-dimensional response variables on covariates that are collected from networks. Although in the current literature considerable attention has been given to methods of learning network topology, little work has been done in the regression analysis, a methodology that plays a central role in studying response-covariate relationships. Because data from a network are correlated across nodes, in order to achieve high statistical efficiency we aim to address an analytic challenge concerning the need of incorporating appropriate dependence structures in parameter estimation and inference.

Networked data have more complex dependence mechanisms than what conventional covariance or correlation matrices may describe. For example, dependence symmetry may be invalid among nodes, and strength of dependence may not be explicitly modelled due to the lack of legitimate distance function between nodes. Our motivating example comes from one of our collaborative projects with scientists in the Center for Human Growth and Development. The scientific objective of the project is to evaluate whether or not, and if so how, iron deficiency affects auditory recognition memory for infants. Infant's memory capability is measured by electrical activities of the brain during a period of 2000 milliseconds using electroencephalography (EEG) net with 64-channel sensors on the scalp (Figure 4.1). The data collection occurs at two time points: when an infant hears his/her mother's voice and when hears a stranger's voice. At each time point, three event-related potentials (ERPs), i.e. P2, P750 and late slow wave (LSW), are reported after the standard data processing. These three ERPs are widely used as primary outcomes of auditory recognition memory (*Mai et al.*, 2012; *Siddappa et al.*, 2004). In this chapter, we consider only the outcome LSW for motivation. Clearly, LSW measurements from 64 electrodes on an infant are correlated in the EEG-net, and such correlation is highly clustered according to subregions of memory functionality. According to our collaborators, cor-

Figure 4.1: Layout of the EGI 64-channel sensor net with 6 outlined clusters of nodes related to auditory recognition memory and 1 additional cluster of the remaining nodes.



relations of LSW measurements are not necessarily symmetric over the 64 nodes. The standard analysis of the data using spatial ANOVA mixed-effects model (*Fields and Kuperberg*, 2012; *Gevins and Smith*, 2000) assumed implicitly symmetric exchangeable correlations among 64 nodes for the LSW data, and failed to detect significant association of iron deficiency on LSW.

To improve the standard analysis our new idea is to recognize the EEG-net as a network, in which we intend to develop a flexible dependence model that can better reflect the underlying relationships among the electrodes; for example, to allow clustered and asymmetric dependence relationships. In particular, we develop a novel strategy to combine two sources of knowledge regarding the network topology in estimation and inference; one source is from expertise of our collaborators regarding the established or prior knowledge about subregions of memory functionality, and the other is the learned dependencies using a statistical method from the available

data at hand. Some of popular statistical methods useful to learn sparse conditional dependence structures of network include sparse partial correlation (*Peng et al.*, 2009) (R package *space*), graphical LASSO (*Yuan and Lin*, 2007) (R package *glasso*), neighborhood selection (*Meinshausen and Buehlmann*, 2006), nonparanormal (*Liu et al.*, 2012) (R package *huge*), among others.

In this chapter we consider the marginal regression model for networked data, because such model has great flexibility on allowing various forms of dependence structures among nodes and its ease on handling categorical outcomes. In contrast, generalized linear mixed effects models for binary data are computationally intricate and may become prohibited when the number of random effects is large. For the estimation of regression coefficients in the marginal model, both generalized estimating equation (GEE) (*Liang and Zeger*, 1986) and quadratic inference function (QIF) (*Qu et al.*, 2000) have been extensively studied in the literature. However, these two methods cannot be directly applied to deal with networked data because of the challenge on incorporating network dependence structures of potentially high dimension. One desirable method to fit the marginal model under unstructured correlation is *Qu and Lindsay* (2003)'s adaptive estimating equation method, which does not require the inverse of correlation matrix. A disadvantage of using unstructured correlation in *Qu and Lindsay* (2003)'s adaptive QIF or GEE is the involvement of a large number of nuisance parameters in the estimation, leading to some potential loss of estimation efficiency and numerical instability. Many authors have advocated the importance of incorporating proper correlation structures in GEE or QIF to achieve desirable estimation efficiency; see for example, *Pan* (2001), *Qu et al.* (2008), *Wang and Carey* (2003) and *Zhou and Qu* (2012).

Our strategy to combine two sources of network topology follows *Stein* (1956)'s linear shrinkage estimation, which was later extensively discussed by *Ledoit and Wolf* (2004) in the context of covariance matrix estimation. We propose to shrink an

unstructured covariance matrix towards a prior network structure (or a target structure) represented by an adjacency matrix with "0" elements for no connection and "1" elements for connection between nodes. Following *Hansen* (1982) we construct an over-identified estimating function, in which a shrinkage tuning parameter is involved and determined by minimizing the inverse of Godambe information to achieve desirable estimation efficiency. As a result, our estimation method will allocate larger weights to more relevant correlation structures and to downweight others. It is worth noting that the process of tuning does not affect estimation consistency nor asymptotic normality but gains efficiency when it is done properly.

This chapter is organized as follows: Section 4.2 briefly describes the QIF method and the adaptive estimating equation method. Section 4.3 introduces the hybrid quadratic inference functions (HQIF) for estimating regression parameters and selecting shrinkage coefficient for networked data, in which the large sample properties are discussed. Section 4.4 presents simulation studies to compare our HQIF method with the popular GEE and the conventional QIF methods. ERPs data is analyzed using the proposed method in Section 4.5, followed by a discussion in Section 4.6. Some technical details are listed in the Appendix D and E.

4.2 Framework

4.2.1 Estimating functions

Consider data arising from a network. Suppose that the response variable y_{ij} and the associated p-dimensional covariate x_{ij} are measured at node (or vertex) j for subject i, $j = 1, \ldots, m$ and $i = 1, \ldots, n$. Let $y_i = (y_{i1}, \ldots, y_{im})^T$, $m \times p$ matrix $x_i = (x_{i1}, \ldots, x_{im})^T$, and (y_i, x_i), $i = 1, \ldots, n$ are i.i.d. samples from n subjects. To perform a regression analysis of the networked data, we adopt the population-average model framework where the mean model is specified by $\mu_{ij} = E(y_{ij}|x_{ij}) = \mu(x_{ij}^T\beta)$,

with $\mu(\cdot)$ being a known link function, β being a p-dimensional parameter vector of interest, and $\mu_i = (\mu_{i1}, \ldots, \mu_{im})^T$.

To proceed with the quasi-likelihood approach to estimating and making inference on β, according to *Liang and Zeger* (1986), the second moment of y_i is specified by $V_i = A_i^{1/2} R(\alpha) A_i^{1/2}$ with $R(\alpha)$ being a working correlation matrix and A_i being the diagonal matrix of the marginal variances $\text{var}(y_{ij}|x_{ij}) = \phi v(\mu_{ij})$, where $v(\cdot)$ is the variance function and ϕ is the dispersion parameter. The seminal work of generalized estimating equations (GEE) (*Liang and Zeger*, 1986) is to obtain an estimate of β by solving equation $\sum_{i=1}^{n} \dot{\mu}_i^T V_i^{-1}(y_i - \mu_i) = 0$, where $\dot{\mu}_i(\cdot)$ is the gradient vector of $\mu_i(\cdot)$ with respect to β; see for example *Song* (2007)'s Chapters 2 and 5 for more details. Because the number of nodes in a network is fixed, we denote variance $V_i \equiv V$. Under some regularity conditions, the resulting GEE estimator is shown to be consistent and asymptotically normal, but may be of low efficiency if working correlation $R(\alpha)$ does not sufficiently represent the true correlation structure.

A variety of strategies have been proposed to improve the efficiency for the GEE estimator. Among them, the QIF method proposed by *Qu et al.* (2000) is of great popularity. QIF approach is based on an assumption that the inverse of the working correlation matrix, R^{-1}, may be expanded approximately as a linear combination of several basis matrices,

$$R^{-1}(\alpha) = \sum_{k=0}^{K} a_k M_k, \tag{4.1}$$

where M_0 is the identity matrix, and M_k, $k = 1, \ldots, K$, are known symmetric basis matrices with 0 and 1 components, and a_k's are unknown coefficients that may depend on parameter α. Then, the GEE may be written as a linear combination of estimating

functions in the following extended score vector,

$$\bar{f}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\beta) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \dot{\mu}_i^T A_i^{-1}(y_i - \mu_i) \\ \dot{\mu}_i^T A_i^{-1/2} M_1 A_i^{-1/2}(y_i - \mu_i) \\ \vdots \\ \dot{\mu}_i^T A_i^{-1/2} M_K A_i^{-1/2}(y_i - \mu_i) \end{pmatrix}, \tag{4.2}$$

where the dimension of $\bar{f}_n(\beta)$ is $p(K+1)$. Unlike the GEE, the QIF does not require estimate nuisance parameter α. Because $\bar{f}_n(\beta)$ is an over-identified score vector, β can not be solved from $\bar{f}_n(\beta) = 0$. Instead, the QIF method is to minimize a quadratic objective function of the following form, similar to *Hansen* (1982)'s idea of generalized method of moments,

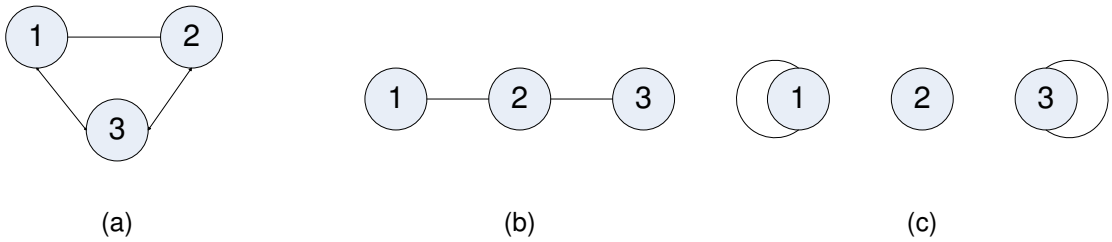
$$n\bar{f}_n^T(\beta)\Gamma^{-1}(\beta)\bar{f}_n(\beta), \tag{4.3}$$

where the optimal weighting matrix is $\Gamma(\beta) = \text{var}\{f_i(\beta)\}$, which may be consistently estimated by its sample covariance matrix $\bar{\Gamma}_n = n^{-1}\sum_{i=1}^{n} f_i(\beta)f_i^T(\beta)$. In implementation, we adopt the unique Moore-Penrose generalized inverse in (4.3) to ensure numerical stability, as the matrix $\bar{\Gamma}_n$ may become singular in some cases (*Hu and Song*, 2012).

4.2.2 Graphic interpretation to basis matrices

We now present some geometric insights on the connection between basis matrices and network topology using two popular correlation structures. This discussion helps us to understand in which form the prior expert's knowledge of network topology may enter the estimation procedure. For the ease of discussion, consider a three-dimensional network. First, the exchangeable correlation matrix, according to *Qu et al.* (2000), has two basis matrices in (4.1), namely $M_0 = I$, and M_1 with 0 on the diagonal and 1 elsewhere. The other example is the AR-1 correlation structure

Figure 4.2: Graphic representations of basis matrices M_{comp} (M_1), M_{chain} (M_1^*) and M_2^* for a network of three nodes.



that has three basis matrices in (4.1), including $M_0 = I$, and M_1^* with 1 on the sub-diagonals and 0 elsewhere, and M_2^* with 1 on the two corner components of the diagonal.

All these basis matrices may be regarded as adjacency matrices, and their graphic representations are displayed in Figure 4.2. It is interesting to note that $M_0 = I$ is an adjacency matrix of independence graph with no connectivity among the nodes; basis matrix M_1 in panel (a) from the exchangeable correlation corresponds to a complete graph, and denoted by M_{comp}; the basis matrices for the AR-1 correlation, M_1^* in panel (b) represents a chain graph and denoted by M_{chain}; and in panel (c) M_2^* indicates both beginning node and ending node as absorbing nodes in a chain graph. Such graphic representation about between-node connectivity is a typical form of network topology knowledge available from subject-matter scientists, or from a learned network derived by the inverse of correlation matrix using training data or pilot study data. More importantly, the QIF theory has demonstrated the feasibility of incorporating adjacency matrices in the estimation and inference via eqn. (4.2) for the parameters in regression models. The key insight here is that each nonzero off-diagonal element in the adjacency matrix (or basis matrix) corresponds to an edge in a graphic model that describes the existence of conditional dependence between two nodes given the other nodes. Since no numeric value of connection strength is available in an adjacency matrix, it is particularly suitable to represent certain prior

knowledge about a network topology. In the case of exchangeable correlation matrix, the complete network adjacency matrix M_{comp} is regarded as being sufficient, since the inverse of the correlation matrix, $R^{-1}(\alpha)$, can be fully represented by basis matrices I and M_{comp}. In the case of AR-1, the chain network adjacency matrix M_{chain} is partially sufficient, since it only captures the conditional dependence among nodes without self-connectivity of the beginning and ending nodes.

4.2.3 Data-driven network topology

Note that the QIF method is easy to be generalized for networked data analysis as long as the adjacency matrices can be constructed in a reasonable manner. In practice, however, the underlying graphic structures from the networked data are so complex that simple structures, such as complete graph in Fig. 4.2(a) and chain graph in Fig. 4.2(b), are not sufficient. On the other hand, using the available data we can establish some data-driven knowledge via, for example, an unstructured dependency in which all variances and covariances are estimated. A drawback of this approach is that in a high-dimensional network the inverse of estimated covariance could be computationally unstable or even prohibited by the standard software. One solution given by *Qu and Lindsay* (2003) is the so-called adaptive procedure that requires only the estimation of covariance matrix. It follows from Cayley-Hamilton theorem (*Bhatia*, 1997) that for an $m \times m$ positive-definite matrix we have

$$V^{-1} = \frac{(-1)^{(m-1)}}{det(V)}(c_1 I + c_2 V + \cdots + c_{m-1} V^{m-2} + V^{m-1}). \tag{4.4}$$

Consequently, the optimal weight matrix $V^{-1}\dot{\mu}$ for a basic estimating function $s = y - \mu(\beta)$ lies in the space spanned by the columns of $\{\dot{\mu}, V\dot{\mu}, \ldots, V^{m-1}\dot{\mu}\}$. For the sake of parsimony, *Qu and Lindsay* (2003) suggested include only the gradient direction generated by the first two columns $\{\dot{\mu}, V\dot{\mu}\}$ in (4.4). This gives the following extended

score vector,

$$\bar{h}_n(\beta) = (\bar{h}_n^{(1)}, \bar{h}_n^{(2)})^T = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \dot{\mu}_i^T(y_i - \mu_i) \\ \dot{\mu}_i^T V(y_i - \mu_i) \end{pmatrix}, \tag{4.5}$$

where V is consistently estimated by $\widehat{V} = \frac{1}{n}\sum_{i=1}^{n} s_i s_i^T$ with $s_i = y_i - \mu_i(\beta)$. Clearly, $\bar{h}_n(\beta)$ in (4.5) does not require the availability of basis matrices given by the expansion in (4.1). However, the number of parameters in matrix V to be estimated is large, especially in the case of large complex networks, and thus "overfitting" may occur in the determination of network dependence structure. Therefore, it seems critical to regularize the covariance matrix estimation, so that the resulting estimated dependencies would balance parsimony and quality of fit to improve statistical power.

4.3 New Methodology

4.3.1 Hybrid quadratic inference function

Inspired by the idea of shrinkage estimation introduced by *Stein* (1956), our regularization procedure considers to shrink estimation of covariance V toward a known prior structure H (an expert's given adjacency matrix). We propose to build up the new extended score $\bar{g}_n$

$$\bar{g}_n(\beta|\gamma) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta|\gamma) = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} \dot{\mu}_i^T A_i^{-1}(y_i - \mu_i) \\ \dot{\mu}_i^T \left\{\gamma A_i^{-1/2} H A_i^{-1/2} + (1-\gamma)V\right\}(y_i - \mu_i) \end{bmatrix}, \tag{4.6}$$

where $\gamma \in [0, 1]$ denotes the shrinkage intensity coefficient. The second component in (4.6) is deemed an improvement in estimation efficiency. Let $U_i(\gamma) = \gamma A_i^{-1/2} H A_i^{-1/2} + (1-\gamma)V$, and $U_i(\gamma)$ is called a linear shrinkage estimator of V (*Ledoit and Wolf*, 2004). Note that for $\gamma = 1$ the shrinkage estimator favors fully the target H, whereas $\gamma = 0$ reduces to the unrestricted covariance V. The key feature of this approach is

that it offers a systematic way to obtain a regularized dependence structure, which outperforms an individual structure $A_i^{-1/2}HA_i^{-1/2}$ or V in terms of numerical stability and statistical efficiency in the estimation of regression parameter β.

Equivalently, the extended score $\bar{g}_n$ in (4.6) can be rewritten as

$$\bar{g}_n(\beta|\gamma) = \frac{\gamma}{n}\sum_{i=1}^{n}\begin{pmatrix} \dot{\mu}_i^T A_i^{-1}(y_i-\mu_i) \\ \dot{\mu}_i^T A_i^{-1/2}HA_i^{-1/2}(y_i-\mu_i) \end{pmatrix} + \frac{(1-\gamma)}{n}\sum_{i=1}^{n}\begin{pmatrix} \dot{\mu}_i^T A_i^{-1}(y_i-\mu_i) \\ \dot{\mu}_i^T V(y_i-\mu_i) \end{pmatrix}. \tag{4.7}$$

Hence, $\bar{g}_n(\beta|\gamma)$ can be viewed as $\gamma\bar{f}_n(\beta|H) + (1-\gamma)\bar{h}_n(\beta|V)$, where γ describes the relative weighting of importance given to $\bar{f}_n$ versus $\bar{h}_n$. For convenience, we call $\bar{g}_n(\beta|\gamma)$ in (4.7) the hybrid extended score vector, which is based on unbiased estimating functions. Note that $\bar{f}_n(\beta|H)$ can produce poor results if the target network structure H is noninformative and far from the truth; similarly, $\bar{h}_n(\beta|V)$ may lose efficiency if certain prior dependency structure is known. Therefore, the proposed $\bar{g}_n$ by allocating higher weights to more relevant extend score vectors has potential promise to improve both computational performance and statistical properties in estimation and inference for β.

Consequently, given shrinkage coefficient γ, we can estimate β by minimizing the following hybrid quadratic inference function (HQIF) Q_n

$$Q_n(\beta|\gamma) = n\bar{g}_n^T(\beta|\gamma)\Gamma^{-1}(\beta|\gamma)\bar{g}_n(\beta|\gamma), \tag{4.8}$$

where Γ can be consistently estimated by $\bar{\Gamma}_n = n^{-1}\sum_{i=1}^{n} g_i(\beta|\gamma)g_i^T(\beta|\gamma)$. Since the estimator of β depends on the choice of shrinkage coefficient γ, it is denoted by $\widehat{\beta}(\gamma)$ in the rest of the chapter.

4.3.2 Asymptotic properties

According to *Hansen* (1982)'s theory of generalized method of moments, under some regularity conditions (*Hansen*, 1982; *Harris et al.*, 1999), it is known that the GMM estimator of β is not only consistent but also asymptotically normally distributed. With a known target structure H, and a fixed shrinkage coefficient γ, these large-sample properties remain valid for the proposed HQIF estimator obtained by minimizing the HQIF (4.8). They are, $\widehat{\beta}(\gamma) \xrightarrow{p} \beta_0$, as $n \to \infty$; and

$$\sqrt{n}(\widehat{\beta}(\gamma) - \beta_0) \xrightarrow{d} N(0, J^{-1}(\beta_0|\gamma)), \text{ as } n \to \infty, \tag{4.9}$$

where $J(\beta_0|\gamma) = G^T(\beta_0|\gamma)\Gamma^{-1}(\beta_0|\gamma)G(\beta_0|\gamma)$, having $\bar{\Gamma}_n(\widehat{\beta}|\gamma) \xrightarrow{p} \Gamma(\beta_0|\gamma)$ and $\dot{\bar{g}}_n(\widehat{\beta}|\gamma) \xrightarrow{p} G(\beta_0|\gamma)$, is the Godambe information of $g_i(\beta_0|\gamma)$. Note that the hybrid extended score vector $g_i(\beta_0|\gamma)$ is constructed on the basis of a known target structure H, so $\widehat{\beta}(\gamma)$ and $J(\beta_0|\gamma)$ depend not only on γ but also on H, and for the notational convenience its dependence on H is not indexed explicitly in the rest of this chapter, unless necessary.

In addition to the above large sample properties, the asymptotic χ^2-distribution of the QIF (*Qu and Lindsay*, 2003; *Qu et al.*, 2000) can be easily extended to the HQIF method with little effort. That is, the hybrid quadratic score-type statistic $\widehat{Q}_n(\widehat{\beta}(\gamma)|\gamma) \xrightarrow{d} \chi^2_{\text{rank}\{\Gamma(\beta_0|\gamma)\}-p}$, which can be used to test for goodness-of-fit under the null hypothesis $H_0 : E(\bar{g}_n) = 0$. Furthermore, a score-type test for a nested model can also be derived. Consider a partition consisting of parameter of interest β_A and other parameter β_B, say $\beta = (\beta_A, \beta_B)$. To test the null hypothesis $H_0 : \beta_A = a_0$, the test statistic takes the form $Q_n(a_0, \tilde{\beta}_B(\gamma)|\gamma) - Q_n(\widehat{\beta}_A(\gamma), \widehat{\beta}_B(\gamma)|\gamma) \xrightarrow{d} \chi^2_{\dim(a_0)}$, where $\tilde{\beta}_B = \arg\min_{\beta_B} Q_n(a_0, \beta_B|\gamma)$, and $(\widehat{\beta}_A(\gamma), \widehat{\beta}_B(\gamma)) = \arg\min_{(\beta_A,\ \beta_B)} Q_n(\beta_A, \beta_B|\gamma)$. Note that this asymptotic χ^2 distribution under $H_0 : \beta_A = a_0$ holds for any $\gamma \in [0, 1]$.

4.3.3 Choice of the shrinkage coefficient

We hope to determine a desirable shrinkage coefficient γ that can achieve a balance between two types of network dependence structures under a certain optimality criterion. In this section we propose to select γ by minimizing the trace of the inverse of the Godambe information matrix $J(\beta_0|\gamma)$ to optimize estimation efficiency over $\gamma \in [0,1]$. That is,

$$\widetilde{\gamma} = \arg\min_{\gamma\in[0,1]} \ \mathrm{tr}\{J^{-1}(\beta_0|\gamma)\}. \tag{4.10}$$

The Godambe information matrix may be consistently estimated by $\widehat{J}(\widehat{\beta}(\gamma)|\gamma) = \dot{\bar{g}}_n^T(\widehat{\beta}(\gamma)|\gamma)\ \bar{\Gamma}_n^{-1}(\widehat{\beta}(\gamma)|\gamma)\dot{\bar{g}}_n(\widehat{\beta}(\gamma)|\gamma)$. Therefore, an estimated norm is $\widehat{\eta}(\gamma) = \mathrm{tr}\left\{\widehat{J}^{-1}(\widehat{\beta}(\gamma)|\gamma)\right\}$, which is the sample counterpart of norm $\eta_0(\gamma) = \mathrm{tr}\left\{J^{-1}(\beta_0|\gamma)\right\}$. Note that $\eta_0(\gamma)$ is continuous on $\gamma \in [0,1]$ and may not be a unimodal function of γ, so there possibly exist multiple shrinkage coefficients that minimize $\eta_0(\gamma)$. In the implementation, greedy searching over a dense grid of γ values is carried out, and let $\gamma_0^* = sup\ \{\widetilde{\gamma}\}$ for all $\widetilde{\gamma}$ minimizing $\eta_0(\gamma)$. Here we pick the largest value γ_0^* to be in more favor of the prior dependency structure H than the unrestricted covariance V. This leads to the unique tuning value for the maximum efficiency.

The following lemma shows that the optimal shrinkage coefficient γ_0^* can be consistently selected when the sample size goes to infinity.

Lemma IV.1. *Let* $S_0 = \{\gamma : \min_{\gamma\in[0,1]} \eta_0(\gamma)\}$ *with* $\eta_0(\gamma) = \mathrm{tr}\left\{J^{-1}(\beta_0|\gamma)\right\}$ *and* $S = \{\gamma : \min_{\gamma\in[0,1]} \widehat{\eta}(\gamma)\}$ *with* $\widehat{\eta}(\gamma) = \mathrm{tr}\left\{\widehat{J}^{-1}(\widehat{\beta}(\gamma)|\gamma)\right\}$. *Suppose that* $|S_0| = |S| < \infty$, *and that both sensitivity matrix* $G(\beta_0|\gamma)$ *and variability matrix* $\Gamma(\beta_0|\gamma)$ *are bounded for* $\gamma \in [0,1]$. *Let* $\gamma_0^* = sup\ \{S_0\}$. *Then* $\widehat{\gamma}^* = sup\ \{S\}$ *is weakly consistent, namely* $\widehat{\gamma}^* \xrightarrow{p} \gamma_0^*$, *as* $n \to \infty$.

The proof for Lemma IV.1 is outlined in the Appendix. Following the standard GMM arguments, we consequently establish the following theorem.

Theorem IV.2. *Under some regularity conditions of GMM (Chapter 1 from Harris et al. (1999)), the regression parameter estimator* $\widehat{\beta}(\widehat{\gamma}^*)$ *at the optimal tuning* $\widehat{\gamma}^* = sup\ \{S\}$ *is asymptotically normal,* $\sqrt{n}(\widehat{\beta}(\widehat{\gamma}^*) - \beta_0) \xrightarrow{d} N(0, J^{-1}(\beta_0|\gamma_0^*))$*, as* $n \to \infty$*.*

Theorem IV.2 indicates that the regression parameter estimator at the optimal shrinkage coefficient $\widehat{\gamma}^*$ is asymptotically normal distributed and more efficient than other estimators obtained under an arbitrary $\gamma \in [0, 1]\backslash S$.

4.4 Simulation Experiment

We conduct simulation studies to evaluate the performance of the proposed HQIF method under an expert prespecified adjacency matrix H^* with an optimal shrinkage coefficient $\widehat{\gamma}^*$, denoted by HQIF($H = H^*, \gamma = \widehat{\gamma}^*$). We consider cases with both continuous and binary responses. And we first present simulation results for networked continuous data. Similar results are also found in binary outcomes. The efficiency of regression parameter estimation is compared under different network structures: complete network, chain network, and 5-subregion network. Correlation matrices $R(\alpha)$ used in data generation corresponding to the following three types of networks:
(N1) complete network with exchangeable correlation $R_{\text{EX}}(\alpha = 0.7)$, where $H^* = M_{\text{comp}}$ is used because basis matrix M_{comp} gives a natural adjacency matrix of a complete network resembling a subregion of similarly active neuro nodes;
(N2) chain network with AR-1 correlation $R_{\text{AR}}(\alpha = 0.7)$, where $H^* = M_{\text{chain}}$ is used because basis matrix M_{chain} provides a relevant adjacency matrix of a chain network mimicking certain neuro-nerve branches;
(N3) two networks of five subregions with function-specific clusters respectively, $R^a_{\text{CL}} =$ block-diag$\{R_{\text{EX}}(\alpha = 0.7), R_{\text{AR}}(\alpha = 0.6), I(\alpha = 0), R_{\text{EX}}(\alpha = 0.5), R_{\text{AR}}(\alpha = 0.8)\}$, $R^b_{\text{CL}} =$ block-diag$\{R_{\text{EX}}(\alpha = 0.4), R_{\text{AR}}(\alpha = 0.6), I(\alpha = 0), R_{\text{EX}}(\alpha = 0.2), R_{\text{AR}}(\alpha = 0.8)\}$, where H^* is a sparse prior target structure given by $H_{\text{CL}} =$ block-diag$\{0, M_{\text{chain}},$

$0, 0, M_{\text{chain}}\}$ with M_{chain} being the adjacency matrix of chain graph.

For each scenario, 500 replications are carried out to draw summary statistics. For each simulation, the optimal shrinkage coefficient $\widehat{\gamma}^*$ is determined using the grid search method over a range from 0 to 1 with 25 equally spaced points. To illustrate estimation efficiency, we consider the mean squared error, $MSE(\widehat{\beta}) = \frac{1}{500}\sum_{s=1}^{500} \|\widehat{\beta}^{(s)} - \beta_0\|_2^2$, and the total variance $Tvar(\widehat{\beta}) = \frac{1}{500}\sum_{s=1}^{500} \text{tr}\{\widehat{\text{var}}(\widehat{\beta}^{(s)})\}$, where $\widehat{\beta}^{(s)}$ is the estimate from the s-th simulation and β_0 is the true parameter. The relative efficiency is measured under two criteria, namely simulated relative efficiency (SRE) and ratio of variances ($Rvar$). SRE (or $Rvar$) is defined as the ratio of MSE (or $Tvar$) between two methods under comparison. In all comparisons, we choose HQIF($H = H^*, \gamma = 1$) (i.e. only prior structure H^* being used in HQIF) as the reference. In addition, we also investigate the finite sample performance of goodness-of-fit test statistic and score-type test statistic between nested models in both aspects of Type I error and power at significance level 0.05.

4.4.1 Networked continuous data

The continuous response variables are generated from a marginal model: $y_{ij} = x_{ij}^T\beta_0 + \epsilon_{ij}$, where $x_{ij} = (x_{ij}^{(1)}, x_{ij}^{(2)})^T$, $x_{ij}^{(1)}$ and $x_{ij}^{(2)}$ are generated independently from $N(\frac{j}{m}, 1)$ with varying means $\frac{j}{m}$ over m nodes, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^T \sim \text{MVN}_m(0, R(\alpha))$, $\beta_0 = (\beta_0^1, \beta_0^2)^T = (1, 1)^T$, n is the sample size ranging from $n = 50, 100, 500$, and m is the number of vertices. The size of complete network N1 or chain network N2 is set as $m = 10$ to mimic a subregion of the brain network in our data analysis, whereas the network of five subregions N3 has varying numbers of vertices, $m = 50, 100, 150$, and the resulting dimension for each block in R_{CL}^a or R_{CL}^b is set as $\frac{m}{5} \times \frac{m}{5}$.

Table 4.1 summarizes results of estimation efficiency obtained by different estimation methods under the three types of network structures: complete network N1, chain network N2, and 5-subregion network N3 (R_{CL}^a). Table 4.2 summarizes

Table 4.1: Summary results of simulated relative efficiency (SRE) and ratio of variances ($Rvar$) of β over 500 simulations for $E(y_{ij}) = x_{ij}^T\beta_0$, where $\beta_0 = (1,1)^T$. Three network structures used are: complete network N1, chain network N2, and 5-subregion network N3 (R^a_{CL}). For each network, the fully prior-based HQIF($H = H^*, \gamma = 1$) is used as reference with $SRE = 1$ and $Rvar = 1$. HQIF($H = H^*, \gamma = \widehat{\gamma}^*$) denotes the HQIF estimator we are interested with the prior network structure H^* and the optimally selected shrinkage coefficient $\widehat{\gamma}^*$.

		n=50		n=100		n=500	
True Network	Method	SRE	$Rvar$	SRE	$Rvar$	SRE	$Rvar$
Complete	HQIF($H = M_{\text{comp}}, \gamma = \widehat{\gamma}^*$)	0.915	1.161	0.916	1.084	0.996	1.017
$H^* = M_{\text{comp}}$	HQIF($\gamma = 0$)	0.913	1.161	0.915	1.084	0.996	1.017
$m = 10$	HQIF($H = M_{\text{chain}}, \gamma = 1$)	0.980	0.983	0.946	0.971	0.959	0.962
	GEE independence	0.956	0.798	0.889	0.824	0.832	0.842
	GEE unstructured	0.007	0.000	0.351	0.436	1.000	1.007
	GEE oracle($R = R_{\text{True}}$)	1.123	0.913	1.033	0.961	0.998	0.990
Chain	HQIF($H = M_{\text{chain}}, \gamma = \widehat{\gamma}^*$)	0.983	1.025	0.994	1.008	1.001	1.001
$H^* = M_{\text{chain}}$	HQIF($\gamma = 0$)	0.862	0.994	0.893	0.963	0.932	0.929
$m = 10$	HQIF($H = M_{\text{comp}}, \gamma = 1$)	0.808	0.779	0.770	0.782	0.775	0.785
	GEE independence	0.812	0.673	0.732	0.695	0.700	0.712
	GEE unstructured	0.010	0.021	0.006	0.003	1.009	1.009
	GEE oracle($R = R_{\text{True}}$)	1.106	0.914	1.061	0.964	1.016	0.997
5-Subregion(R^a_{CL})	HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$)	1.302	2.202	1.593	1.971	1.623	1.788
$H^* = H_{\text{CL}}$	HQIF($\gamma = 0$)	1.279	2.106	1.528	1.882	1.558	1.727
$m = 50$	HQIF($H = M_{\text{comp}}, \gamma = 1$)	0.989	0.940	0.943	0.948	0.946	0.954
	HQIF($H = M_{\text{chain}}, \gamma = 1$)	1.504	1.521	1.495	1.501	1.478	1.500
	GEE independence	1.020	0.894	0.967	0.922	0.955	0.946
	GEE oracle($R = R_{\text{True}}$)	2.504	2.325	2.772	2.416	2.292	2.513
	HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$)	1.181	2.533	1.458	2.109	1.748	1.865
	HQIF($\gamma = 0$)	1.183	1.906	1.265	1.612	1.348	1.421
$m = 100$	HQIF($H = M_{\text{comp}}, \gamma = 1$)	0.799	0.790	0.834	0.776	0.790	0.793
	HQIF($H = M_{\text{chain}}, \gamma = 1$)	1.048	1.149	1.195	1.119	1.182	1.129
	GEE independence	0.832	0.748	0.863	0.758	0.799	0.787
	GEE oracle($R = R_{\text{True}}$)	2.503	2.173	2.553	2.266	2.537	2.359
	HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$)	1.033	2.117	1.358	1.823	1.409	1.569
	HQIF($\gamma = 0$)	1.147	1.826	1.275	1.547	1.284	1.350
$m = 150$	HQIF($H = M_{\text{comp}}, \gamma = 1$)	0.656	0.649	0.672	0.649	0.659	0.652
	HQIF($H = M_{\text{chain}}, \gamma = 1$)	0.942	0.861	0.905	0.850	0.741	0.855
	GEE independence	0.694	0.614	0.686	0.631	0.654	0.648
	GEE oracle($R = R_{\text{True}}$)	2.443	2.008	2.309	2.069	2.067	2.140

Table 4.2: Summary results of simulated relative efficiency (SRE) and ratio of variances ($Rvar$) of β over 500 simulations for $E(y_{ij}) = x_{ij}^T\beta_0$, where $\beta_0 = (1,1)^T$. The network structure of outcomes is a 5-subregion network N3 with R_{CL}^b. For each case, the fully prior-based HQIF($H = H_{CL}, \gamma = 1$) is used as reference with $SRE = 1$ and $Rvar = 1$. HQIF($H = H_{CL}, \gamma = \widehat{\gamma}^*$) denotes the HQIF estimator we are interested with the prior network structure H_{CL} and the optimally selected shrinkage coefficient $\widehat{\gamma}^*$.

5-Subregion		n=50		n=100		n=500	
Network (R_{CL}^b)	Method	SRE	$Rvar$	SRE	$Rvar$	SRE	$Rvar$
$H^* = H_{\text{CL}}$	HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$)	0.985	2.215	1.316	1.882	1.482	1.623
	HQIF($\gamma = 0$)	0.924	2.061	1.268	1.754	1.354	1.517
$m = 50$	HQIF($H = M_{\text{comp}}, \gamma = 1$)	0.981	0.945	0.957	0.957	0.946	0.964
	HQIF($H = M_{\text{chain}}, \gamma = 1$)	1.273	1.308	1.306	1.299	1.259	1.297
	GEE independence	1.014	0.900	0.982	0.932	0.954	0.956
	GEE oracle($R = R_{\text{True}}$)	1.815	1.692	1.970	1.762	1.648	1.834
	HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$)	0.837	2.652	1.127	2.036	1.496	1.621
	HQIF($\gamma = 0$)	0.932	2.230	1.077	1.735	1.308	1.394
$m = 100$	HQIF($H = M_{\text{comp}}, \gamma = 1$)	0.940	0.919	0.943	0.908	0.924	0.925
	HQIF($H = M_{\text{chain}}, \gamma = 1$)	1.076	1.170	1.188	1.146	1.202	1.155
	GEE independence	0.975	0.872	0.978	0.888	0.934	0.919
	GEE oracle($R = R_{\text{True}}$)	1.904	1.655	1.897	1.735	1.963	1.803
	HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$)	0.837	2.347	1.108	1.923	1.297	1.561
	HQIF($\gamma = 0$)	0.841	2.289	1.085	1.756	1.200	1.350
$m = 150$	HQIF($H = M_{\text{comp}}, \gamma = 1$)	0.842	0.822	0.844	0.826	0.839	0.830
	HQIF($H = M_{\text{chain}}, \gamma = 1$)	1.056	0.983	1.027	0.974	0.865	0.978
	GEE independence	0.880	0.780	0.861	0.804	0.832	0.825
	GEE oracle($R = R_{\text{True}}$)	1.961	1.596	1.846	1.646	1.645	1.701

results of estimation efficiency obtained by different estimation methods under the 5-subregion network N3 (R^b_{CL}). Here we focus on the comparison of HQIF estimators obtained under HQIF($H = H^*, \gamma = \widehat{\gamma}^*$) to other estimates obtained respectively from the complete structure HQIF($H = M_{\text{comp}}, \gamma = 1$), the chain structure HQIF($H = M_{\text{chain}}, \gamma = 1$), and the fully data-driven structure HQIF($\gamma = 0$). Besides, our method is also compared to GEE estimates under independence correlation representing the independence network, under unstructured correlation, and under the true correlation $R(\alpha)$. The GEE with the true correlation represents the "oracle" case with the true correlation parameter α, because it is of semiparametric efficiency. In the 5-subregion network N3 setting, the GEE unstructured estimation is not provided, because of numerical failure in the case of 100-dimensional network.

From Table 4.1 and Table 4.2, we can see that HQIF($H = H^*, \gamma = \widehat{\gamma}^*$) shows a steady rise in *SRE* and fall in *Rvar* when n increases in all different types of networks. Also, it is not surprising to see that the performance of the GEE unstructured estimator is the worst under moderate sample size ($n = 50$ or $n = 100$), because in this case a large number of correlation parameters need to be estimated. When the true network is the complete graph N1, the *SRE* and *Rvar* of HQIF($H = M_{\text{comp}}, \gamma = \widehat{\gamma}^*$) are very similar to those given by the data-driven HQIF($\gamma = 0$) regardless of sample sizes. When the true network is the chain graph N2, the performance of HQIF($H = M_{\text{chain}}, \gamma = \widehat{\gamma}^*$) becomes closer to that of the GEE oracle when the sample size increases. In the scenario of the 5-subregion graph N3 (R^a_{CL} or R^b_{CL}), it is easy to see that HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$) appears to be the top performer, particularly superior to HQIF($H = H_{\text{CL}}, \gamma = 1$) and HQIF($\gamma = 0$) under $n = 100, 500$. It also works much better than the other working target structures, such as the complete structure by HQIF($H = M_{\text{comp}}, \gamma = 1$), and the independence structure by GEE independence. We also find that HQIF($\gamma = 0$) and HQIF($H = M_{\text{chain}}, \gamma = 1$) may perform better than the HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$) method when $n = 50$. These phenomena could be

expected because estimators have not achieved asymptotic unbiasedness with small sample sizes.

When the network structure is specified as a more realistic scenario of the 5-subregion network N3 with varying network size ($m = 50, 100, 150$), we present the results of optimal shrinkage coefficient selection obtained under R^a_{CL} in Fig. 4.3 and obtained under R^b_{CL} in Fig. 4.4. Specifically, the histograms of the selected optimal shrinkage coefficient $\widehat{\gamma}^*$ for HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$) method show that the percentage of sample counterpart $\widehat{\gamma}^*$ falling near the optimal value γ_0^* increases as the sample size increases. This confirms the selection consistency, $\widehat{\gamma}^* \xrightarrow{p} \gamma_0^*$, as $n \to \infty$ given in Lemma IV.1. Besides, the fourth column in Fig. 4.3 or Fig. 4.4 shows that the target structure H_{CL} tends to be more weighted ($\widehat{\gamma}^* > \frac{1}{2}$) and thus more informative than the unrestricted covariance V with the increase of the network size. In addition, we also present the results of estimation efficiency SRE obtained under R^a_{CL} and R^b_{CL} with sample size $n = 100$ and 500 in Fig. 4.5. We can see that the proposed HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$) outperforms the other approaches, judged by its lower SRE relative to the GEE oracle with $SRE = 1$. When the sample size is 500, the HQIF method utilizing both prior and data-driven information is clearly superior to the other approaches.

To investigate the performance of test statistics given in Section 4.3.2, here we consider the following simulation setup. The full model is given by $y_{ij} = x_{ij}^T\beta_0 + \theta z_i + \epsilon_{ij}$, where z_i is a subject-level variable generated from a Bernoulli distribution with probability 0.5, while x_{ij} and ϵ_{ij} are generated by the same distributions above. The hypothesis of interest is $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Type I error rates are computed with $\theta = 0$, while power is calculated under $\theta = 0.2$. For our HQIF method, the size and power of score-type test are obtained by averaging over 25 candidate shrinkage coefficients in the range from 0 to 1 to remove the influence of γ selection.

Table 4.3 summarizes empirical Type I errors and power of score-type test statis-

tics constructed from the HQIF method at significance level 0.05 over 500 replications. It is clear that the Type I error is well controlled in all cases, and the power increases as the sample size increases. Specifically, when sample size is large ($n = 500$), HQIF($H = H^*, \gamma = 1$) with an expert prespecified prior target H^* performs slightly better than HQIF($\gamma = 0$) and HQIF($H = H^*, \gamma \in [0, 1]$) for the complete network and the chain network. When HQIF($\gamma = 0$) is compared with HQIF($H = H^*, \gamma \in [0, 1]$), we find that their results are only marginally different. For the 5-subregion network N3 (R^a_{CL}), when the sample size is $n = 100$, HQIF($\gamma = 0$) performs slightly better than HQIF($H = H_{\text{CL}}, \gamma = 1$) with a prespecified prior target H_{CL}. When compared with HQIF($H = H_{\text{CL}}, \gamma \in [0, 1]$), the results differ only marginally. In addition, for the case of N3 (R^a_{CL} or R^b_{CL}) with network size $m = 100$, Fig. 4.6 and Fig. 4.7 show QQ-plots of test statistics $Q_n(a_0, \tilde{\beta}_B|\gamma) - Q_n(\widehat{\beta}_A, \widehat{\beta}_B|\gamma)$ and $\widehat{Q}_n(\widehat{\beta}|\gamma)$ based on 500 simulated datasets with sample size $n = 50$ and $n = 500$, respectively. For testing $H_0 : \theta = 0$, it is clear that the plots of the null distribution for the test statistic $Q_n(a_0, \tilde{\beta}_B|\gamma) - Q_n(\widehat{\beta}_A, \widehat{\beta}_B|\gamma)$ indicate satisfactory approximation to the χ^2_1 distribution, even though the sample size is fairly small $n = 50$ in Fig. 4.6. For the goodness of fit test $H_0 : E(\bar{g}_n) = 0$, the QQ-plots of the null distribution for the test statistic $\widehat{Q}_n(\widehat{\beta}|\gamma)$ also closely follows χ^2_3 distribution. These results demonstrate that the null distribution for the score-type testing approach is not sensitive to the choice of the prior network structure H or the shrinkage coefficient γ. However, it is worth to point out that the Wald test statistics, which depend on $\widehat{\beta}$ and var($\widehat{\beta}$), are dependent on the selection of H and γ. This is an advantage of the QIF over the GEE in the hypothesis test.

Table 4.3: Average empirical Type I error rates and power of test statistics at significance level 0.05 over 500 replications. Three network structures used are: complete network N1, chain network N2, and 5-subregion network N3 (R^a_{CL}).

		n=50		n=100		n=500	
Network	HQIF	Size	Power	Size	Power	Size	Power
Complete	$H^* = M_{\mathrm{comp}}, \gamma \in [0,1]$	0.031	0.112	0.063	0.243	0.050	0.764
$m = 10$	$\gamma = 0$	0.030	0.114	0.062	0.242	0.052	0.764
	$H^* = M_{\mathrm{comp}}, \gamma = 1$	0.028	0.100	0.066	0.234	0.046	0.772
Chain	$H^* = M_{\mathrm{chain}}, \gamma \in [0,1]$	0.040	0.154	0.061	0.380	0.051	0.955
$m = 10$	$\gamma = 0$	0.042	0.156	0.062	0.366	0.054	0.952
	$H^* = M_{\mathrm{chain}}, \gamma = 1$	0.038	0.154	0.060	0.396	0.046	0.958
5-Subregion	$H^* = H_{\mathrm{CL}}, \gamma \in [0,1]$	0.049	0.615	0.049	0.945	0.044	1.000
$m = 50$	$\gamma = 0$	0.046	0.642	0.046	0.952	0.048	1.000
	$H^* = H_{\mathrm{CL}}, \gamma = 1$	0.052	0.546	0.056	0.920	0.042	1.000
	$H^* = H_{\mathrm{CL}}, \gamma \in [0,1]$	0.043	0.784	0.055	0.986	0.056	1.000
$m = 100$	$\gamma = 0$	0.044	0.796	0.056	0.992	0.056	1.000
	$H^* = H_{\mathrm{CL}}, \gamma = 1$	0.048	0.656	0.060	0.958	0.048	1.000
	$H^* = H_{\mathrm{CL}}, \gamma \in [0,1]$	0.066	0.878	0.055	0.997	0.068	1.000
$m = 150$	$\gamma = 0$	0.072	0.898	0.052	1.000	0.078	1.000
	$H^* = H_{\mathrm{CL}}, \gamma = 1$	0.062	0.754	0.064	0.986	0.050	1.000

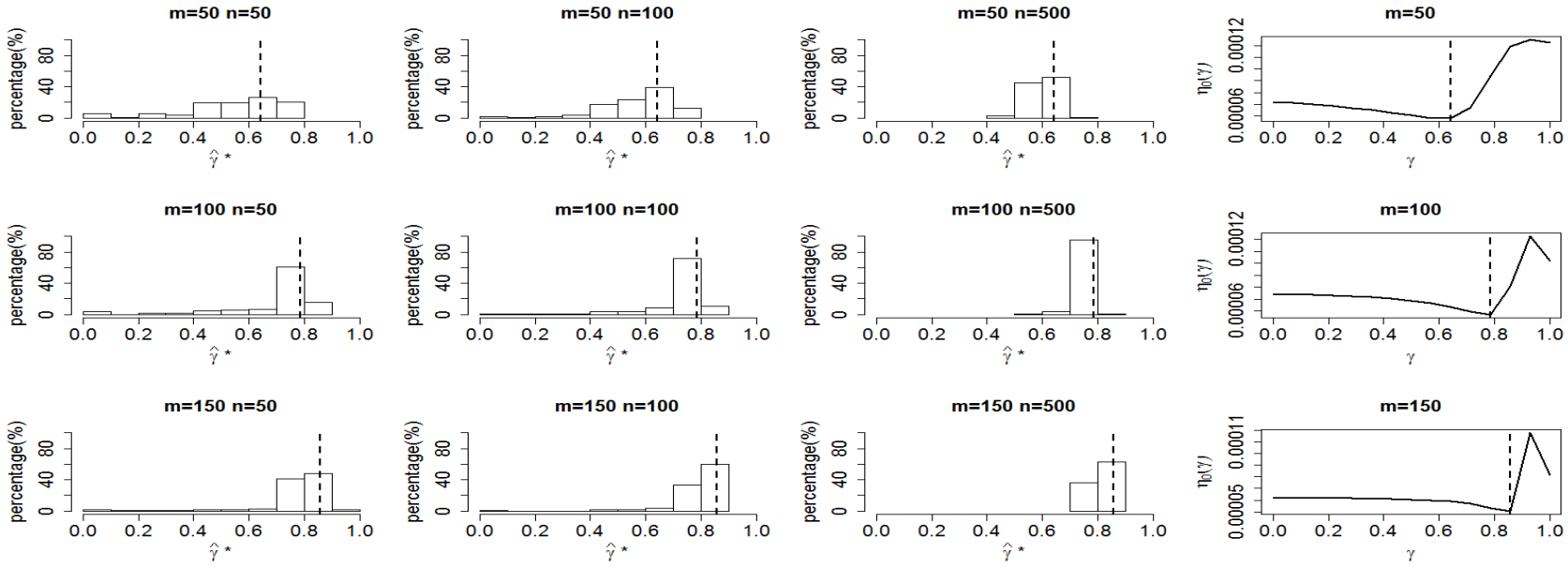


Figure 4.3: The histogram of $\widehat{\gamma}^*$ over 500 simulations and the plot of shrinkage coefficient selection norm $\eta_0(\gamma) = \mathrm{tr}\left\{J^{-1}(\beta_0|\gamma, H^*, R(\alpha))\right\}$ versus γ for the HQIF$(H = H^*, \gamma = \widehat{\gamma}^*)$ method. The network structure of the 5-subregion N3 (R^a_{CL}) is used with $H^* = H_{\mathrm{CL}}$. The first three columns are the histogram of $\widehat{\gamma}^*$ for $m = 50, 100, 150$ and $n = 50, 100, 500$, and the fourth column is the plot of $\eta_0(\gamma)$. In each subplot, the optimal value γ_0^* is given by the vertical line.

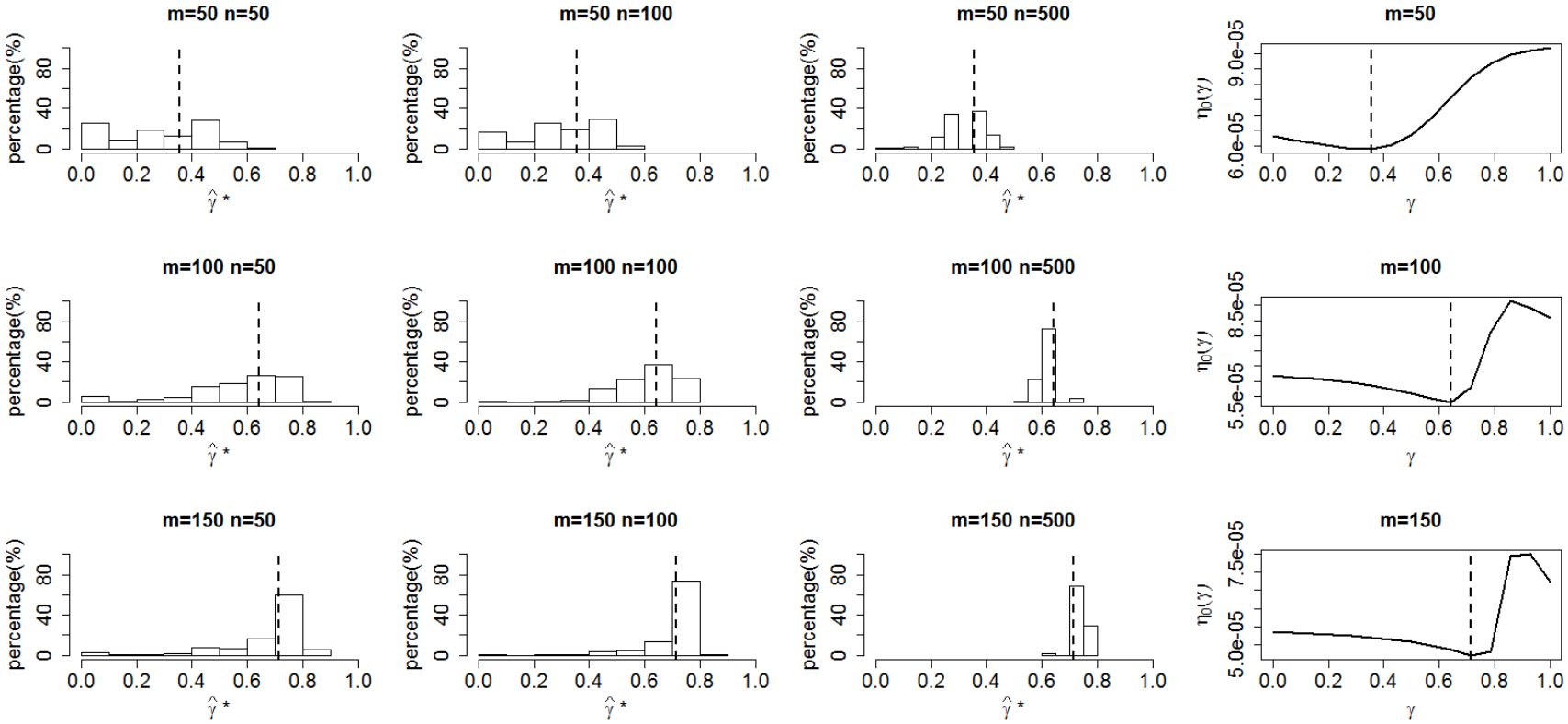


Figure 4.4: The histogram of $\widehat{\gamma}^*$ over 500 simulations and the plot of shrinkage coefficient selection norm $\eta_0(\gamma) = \text{tr}\left\{J^{-1}(\beta_0|\gamma, H^*, R(\alpha))\right\}$ versus γ for the HQIF method. The network structure of the 5-subregion N3 (R^b_{CL}) is used with $H^* = H_{\text{CL}}$. The first three columns are the histogram of $\widehat{\gamma}^*$ for $m = 50, 100, 150$ and $n = 50, 100, 500$, and the fourth column is the plot of $\eta_0(\gamma)$. In each subplot, the optimal value γ_0^* is given by the vertical line.

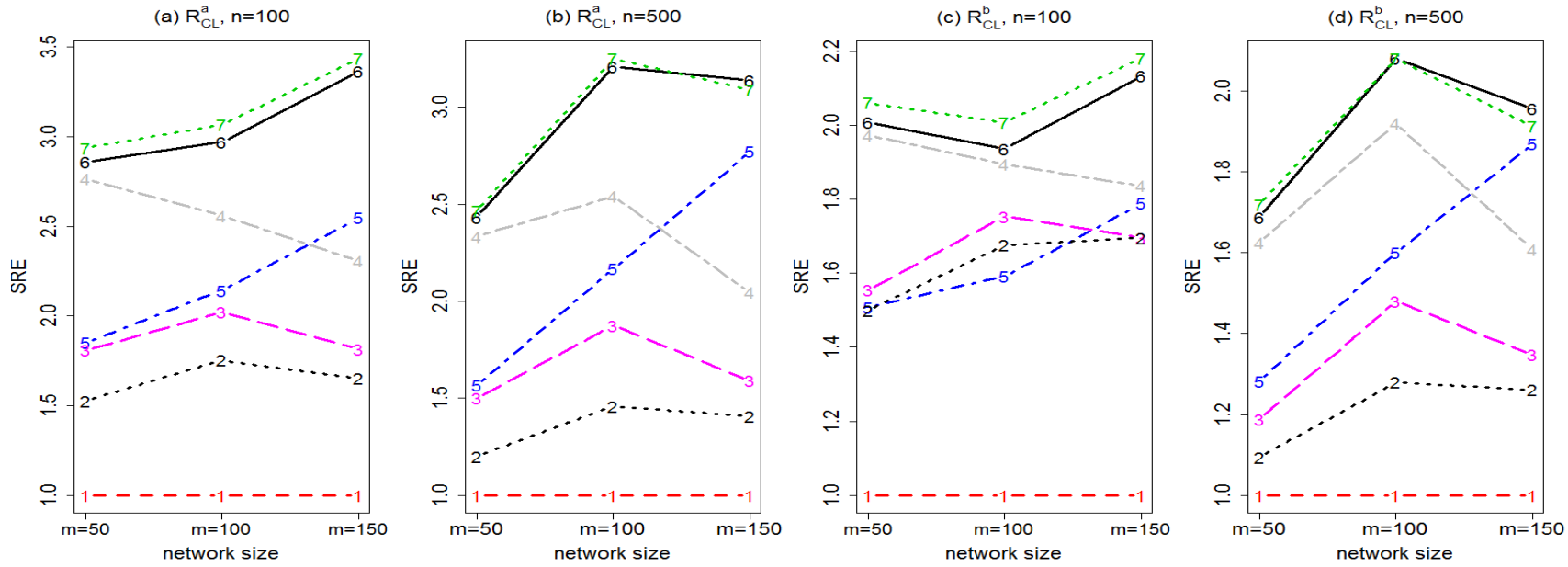


Figure 4.5: *SRE* comparison for continuous data with 5-subregion network N3, where the reference level is GEE oracle, the sample size $n = 100, 500$ and the number of nodes varies in $m = 50, 100, 150$. Plots (a) and (b) are obtained from network dependence structure R^a_{CL}, and plots (c) and (d) are obtained from the structure R^b_{CL}. The index of each line is defined as 1: GEE oracle; 2: HQIF($H = H_{\text{CL}}; \gamma = \widehat{\gamma}^*$); 3: HQIF($\gamma = 0$); 4: HQIF($H = H_{\text{CL}}, \gamma = 1$); 5: HQIF($H = M_{\text{chain}}, \gamma = 1$); 6: GEE independence; 7: HQIF($H = M_{\text{comp}}, \gamma = 1$).

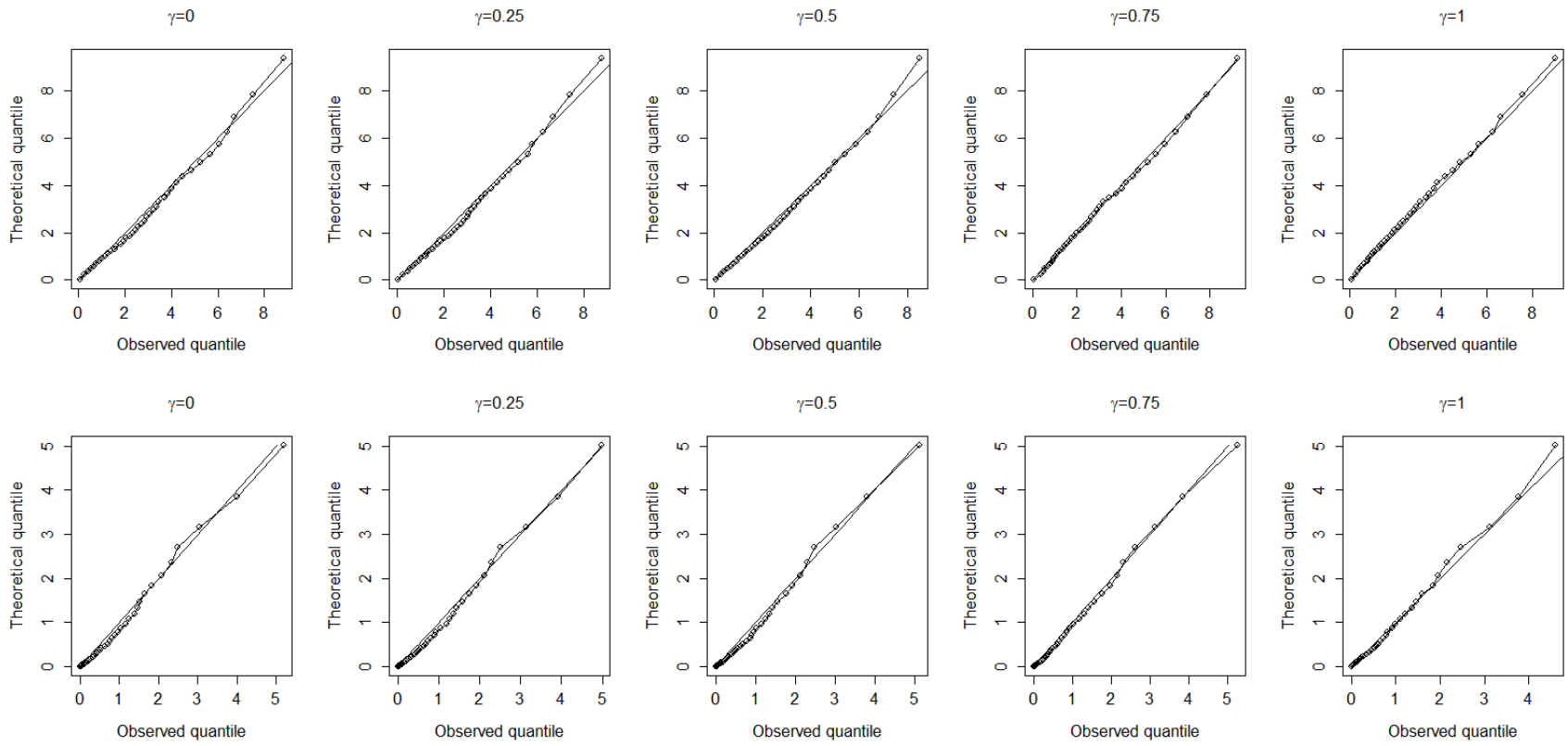


Figure 4.6: QQ-plots of the null distribution for HQIF ($H = H_{\mathrm{CL}}, \gamma = 0, 0.25, 0.5, 0.75, 1$) in 5-subregion network N3 (R^a_{CL}) with $n = 50$. On the top panel: $\widehat{Q}_n(\widehat{\beta}|\gamma)$ relative to χ^2_3 and on the bottom panel $Q_n(a_0, \tilde{\beta}_B|\gamma) - Q_n(\widehat{\beta}_A, \widehat{\beta}_B|\gamma)$ relative to χ^2_1.

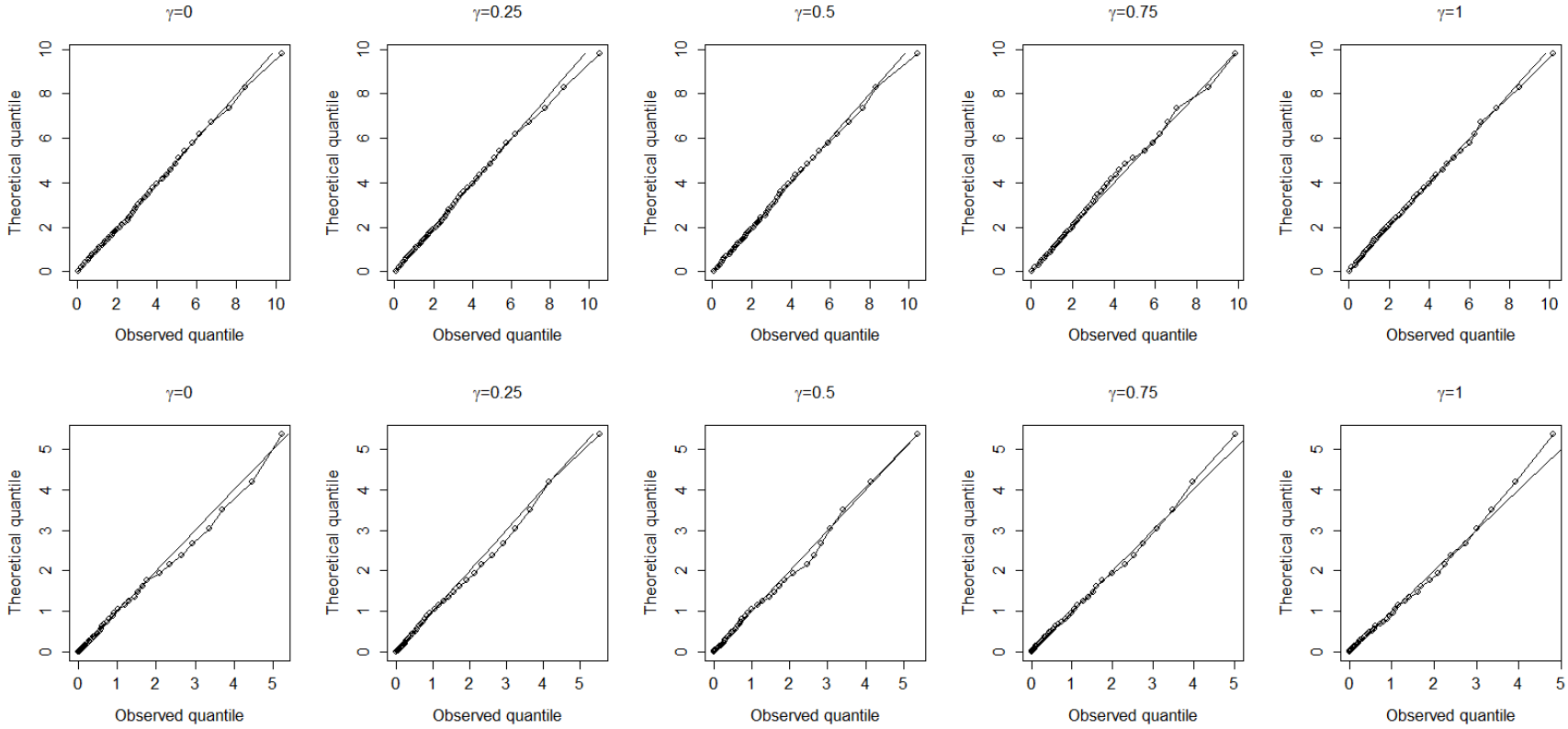


Figure 4.7: QQ-plots of the null distribution for HQIF ($H = H_{\mathrm{CL}}, \gamma = 0, 0.25, 0.5, 0.75, 1$) in 5-subregion network N3 (R^b_{CL}) with $n = 500$. On the top panel: $\widehat{Q}_n(\widehat{\beta}|\gamma)$ relative to χ^2_3 and on the bottom panel $Q_n(a_0, \tilde{\beta}_B|\gamma) - Q_n(\widehat{\beta}_A, \widehat{\beta}_B|\gamma)$ relative to χ^2_1.

4.4.2 Networked binary data

The correlated binary responses in a network are generated using R package *mvtBinaryEP* from the following logistic model: $\text{logit}(\mu_{ij}) = x_{ij}^T \beta_0$, with $x_{ij} = (x_{ij}^{(1)}, x_{ij}^{(2)})^T$. A node-varying covariate is given by $x_{ij}^{(1)} = \frac{j}{m}$, $j = 1, \ldots, m$, and the other covariate $x_{ij}^{(2)}$ is generated independently from $U(0, 1)$; n is the sample size ranging from $n = 50, 100, 500$, and m is the number of vertices for a network. We consider the same three network structures as those given in the beginning of Section 4.4: the complete network N1 (R_{EX}), the chain network N2 (R_{AR}) and the 5-subregion network N3 (R_{CL}^a), which have already been used in the simulation study of continuous data. The size of complete network N1 or chain network N2 is set as $m = 10$ to mimic a subregion of the EEG nodes in our data analysis with $\beta_0 = (\beta_0^1, \beta_0^2)^T = (0.5, 0.2)^T$. The network of five subregions N3 has varying numbers of vertices, $m = 25, 50$, with $\beta_0 = (\beta_0^1, \beta_0^2)^T = (0.5, 0.1)^T$, and the resulting dimension for each block in R_{CL}^a is set as $\frac{m}{5} \times \frac{m}{5}$. The prior adjacency matrix H^* for the complete network N1 and the chain network N2 are M_{comp} and M_{chain}, respectively. In addition, H^* for the 5-subregion network N3 is given by $H_{\text{CL}} = \text{block-diag}\{M_{comp}, M_{chain}, 0, M_{comp}, M_{chain}\}$, where M_{comp} and M_{chain} are the adjacency matrices of complete and chain graphs, respectively.

Table 4.4: Summary results of simulated relative efficiency (SRE) and ratio of variances ($Rvar$) of β over 500 simulations for $\text{logit}(\mu_{ij}) = x_{ij}^T\beta_0$. Three network structures used are: complete network N1, chain network N2, and 5-subregion network N3 (R^a_{CL}). For each network, the fully prior-based HQIF($H = H^*, \gamma = 1$) is used as reference with $SRE = 1$ and $Rvar = 1$. HQIF($H = H^*, \gamma = \widehat{\gamma}^*$) denotes the HQIF estimator we are interested with the prior network structure H^* and the optimally selected shrinkage coefficient $\widehat{\gamma}^*$.

		n=50		n=100		n=500	
Network	Method	SRE	$Rvar$	SRE	$Rvar$	SRE	$Rvar$
Complete	HQIF($H = M_{\text{comp}}, \gamma = \widehat{\gamma}^*$)	0.978	1.078	0.976	1.038	0.992	1.005
$H^* = M_{\text{comp}}$	HQIF($\gamma = 0$)	0.869	1.071	0.945	1.038	0.964	1.005
$m = 10$	HQIF($H = M_{\text{chain}}, \gamma = 1$)	0.945	0.928	0.932	0.917	0.887	0.902
	GEE independence	0.754	0.611	0.695	0.615	0.629	0.621
	GEE unstructured	0.000	0.000	0.891	1.109	0.968	1.020
	GEE oracle($R = R_{\text{True}}$)	1.187	0.970	1.078	0.983	1.015	1.000
Chain	HQIF($H = M_{\text{chain}}, \gamma = \widehat{\gamma}^*$)	1.002	1.022	0.999	1.006	1.000	1.001
$H^* = M_{\text{chain}}$	HQIF($\gamma = 0$)	0.883	1.010	0.915	0.975	0.939	0.950
$m = 10$	HQIF($H = M_{\text{comp}}, \gamma = 1$)	0.778	0.804	0.834	0.811	0.799	0.818
	GEE independence	0.800	0.717	0.794	0.729	0.743	0.745
	GEE unstructured	0.000	0.000	0.000	0.000	0.996	1.019
	GEE oracle($R = R_{\text{True}}$)	1.102	0.961	1.047	0.976	1.012	1.003
5-Subregion	HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$)	0.889	1.193	0.932	1.113	1.028	1.046
$H^* = H_{\text{CL}}$	HQIF($\gamma = 0$)	0.851	1.198	0.939	1.114	1.028	1.046
$m = 25$	HQIF($H = M_{\text{comp}}, \gamma = 1$)	0.795	0.836	0.899	0.832	0.832	0.836
	HQIF($H = M_{\text{chain}}, \gamma = 1$)	0.887	0.931	0.962	0.930	0.898	0.932
	GEE independence	0.849	0.777	0.915	0.790	0.826	0.809
	GEE oracle($R = R_{\text{True}}$)	1.331	1.137	1.280	1.165	1.228	1.203
	HQIF($H = H_{\text{CL}}, \gamma = \widehat{\gamma}^*$)	0.930	1.226	0.998	1.116	1.051	1.049
	HQIF($\gamma = 0$)	0.929	1.227	0.998	1.116	1.051	1.049
$m = 50$	HQIF($H = M_{\text{comp}}, \gamma = 1$)	0.809	0.838	0.843	0.835	0.843	0.833
	HQIF($H = M_{\text{chain}}, \gamma = 1$)	0.751	0.927	0.909	0.918	0.937	0.921
	GEE independence	0.908	0.764	0.863	0.778	0.802	0.791
	GEE oracle($R = R_{\text{True}}$)	1.534	1.285	1.314	1.314	1.305	1.344

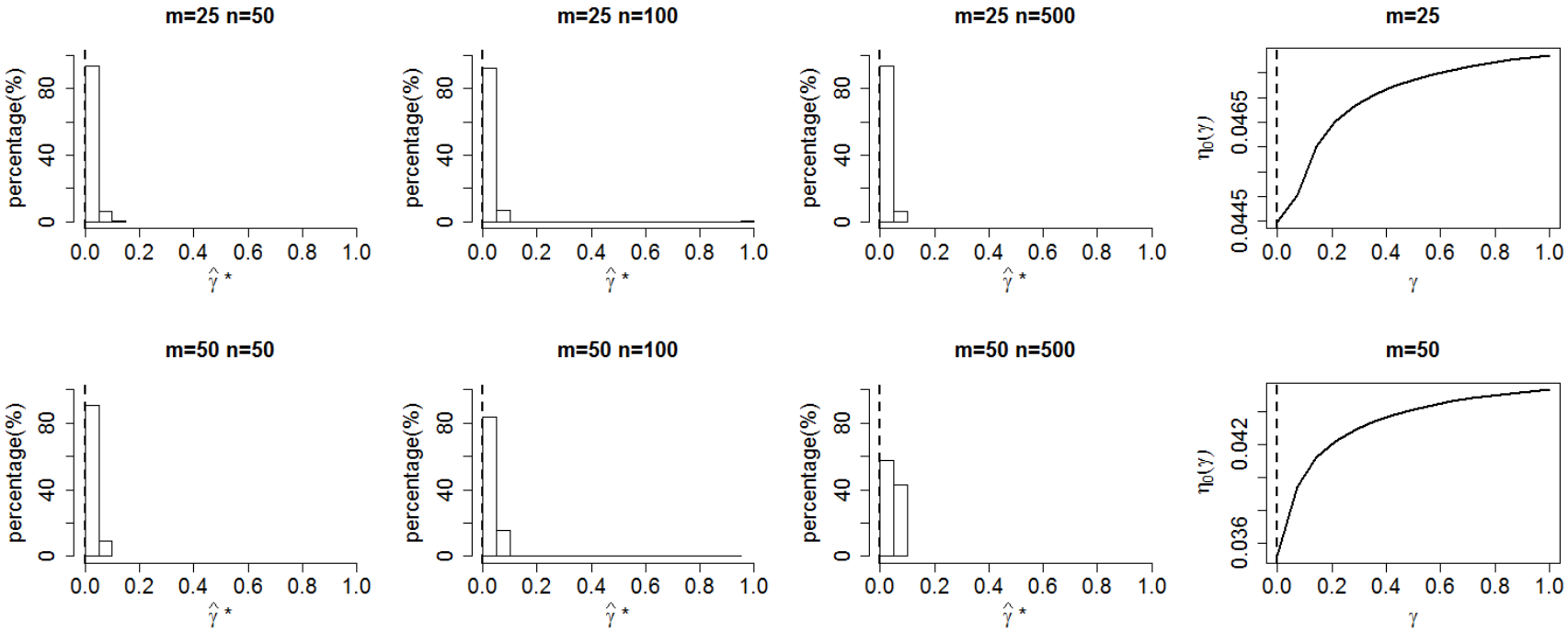


Figure 4.8: The histogram of $\widehat{\gamma}^*$ over 500 simulations and the plot of shrinkage coefficient selection norm $\eta_0(\gamma) = \mathrm{tr}\left\{J^{-1}(\beta_0|\gamma, H^*, R(\alpha))\right\}$ versus γ for the HQIF$(H = H^*, \gamma = \widehat{\gamma}^*)$ method. The network structure of the 5-subregion N3 (R^a_{CL}) is used with $H^* = H_{\mathrm{CL}}$. The first three columns are the histogram of $\widehat{\gamma}^*$ for $m = 25, 50$ and $n = 50, 100, 500$, and the fourth column is the plot of $\eta_0(\gamma)$. In each subplot, the optimal value γ_0^* is given by the vertical line.

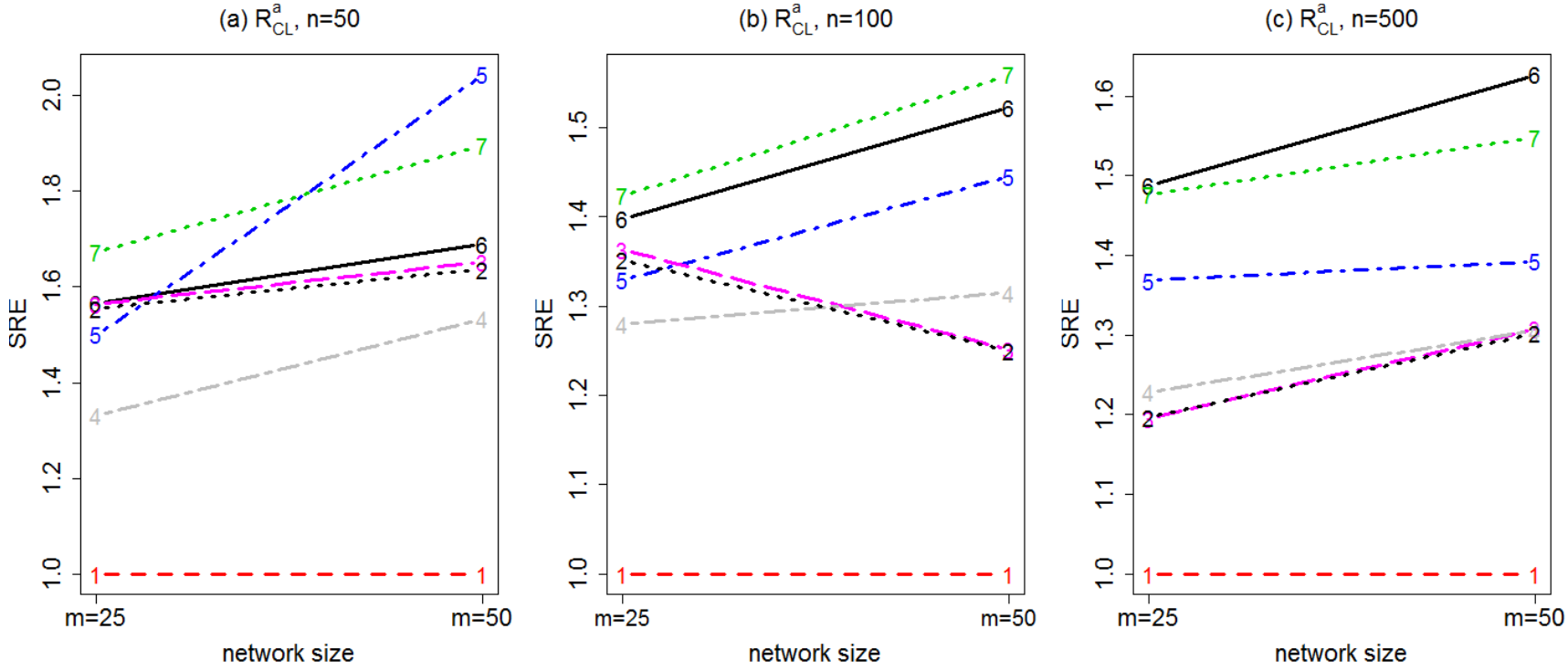


Figure 4.9: SRE comparison for binary data with 5-subregion network N3 (R^a_{CL}), where the reference level is GEE oracle. The sample size varies in $n = 50, 100, 500$ and the number of nodes varies in $m = 25, 50$. The index of each line is defined as 1: GEE oracle; 2: HQIF($H = H_{\mathrm{CL}}, \gamma = \widehat{\gamma}^*$); 3: HQIF($\gamma = 0$); 4: HQIF($H = H_{\mathrm{CL}}, \gamma = 1$); 5: HQIF($H = M_{\mathrm{chain}}, \gamma = 1$); 6: GEE independence; 7: HQIF($H = M_{\mathrm{comp}}, \gamma = 1$).

Table 4.5: Average empirical Type I error rates and power of test statistics at significance level 0.05 over 500 replications. The network structure used here is the 5-subregion network N3 (R^a_{CL}).

5-Subregion Network	HQIF $H^* = H_{\text{CL}}$	n=50		n=100		n=500	
		Size	Power	Size	Power	Size	Power
$m = 25$	$\gamma \in [0, 1]$	0.020	0.150	0.030	0.295	0.044	0.897
	$\gamma = 0$	0.025	0.170	0.025	0.315	0.035	0.930
	$\gamma = 1$	0.020	0.155	0.030	0.290	0.045	0.890
$m = 50$	$\gamma \in [0, 1]$	0.034	0.122	0.022	0.294	0.080	0.879
	$\gamma = 0$	0.025	0.120	0.035	0.300	0.080	0.895
	$\gamma = 1$	0.004	0.125	0.025	0.290	0.085	0.875

To investigate the performance of test statistics given in Section 4.3.2, here we consider the same hypothesis of subject-level effect as that considered in the first simulation study under $H_0 : \theta = 0$, in a population-average logistic model $\text{logit}(\mu_{ij}) = x_{ij}^T\beta + \theta z_i$ for 5-subregion network N3 (R^a_{CL}). Type I errors are computed with $\theta = 0$, while power is calculated with $\theta = 0.2$.

Tables 4.4-4.5 and Figures 4.8-4.10 present results summarised over 500 replications. Most of conclusions remain similar to those drawn in the case of continuous outcomes. However, for the 5-subregion network N3 (R^a_{CL}), the prior target H_{CL} does not appear to be informative compared to the sample covariance regardless of the sample size. This once again indicates the importance of adding a proper and relevant prior H matrix in the HQIF.

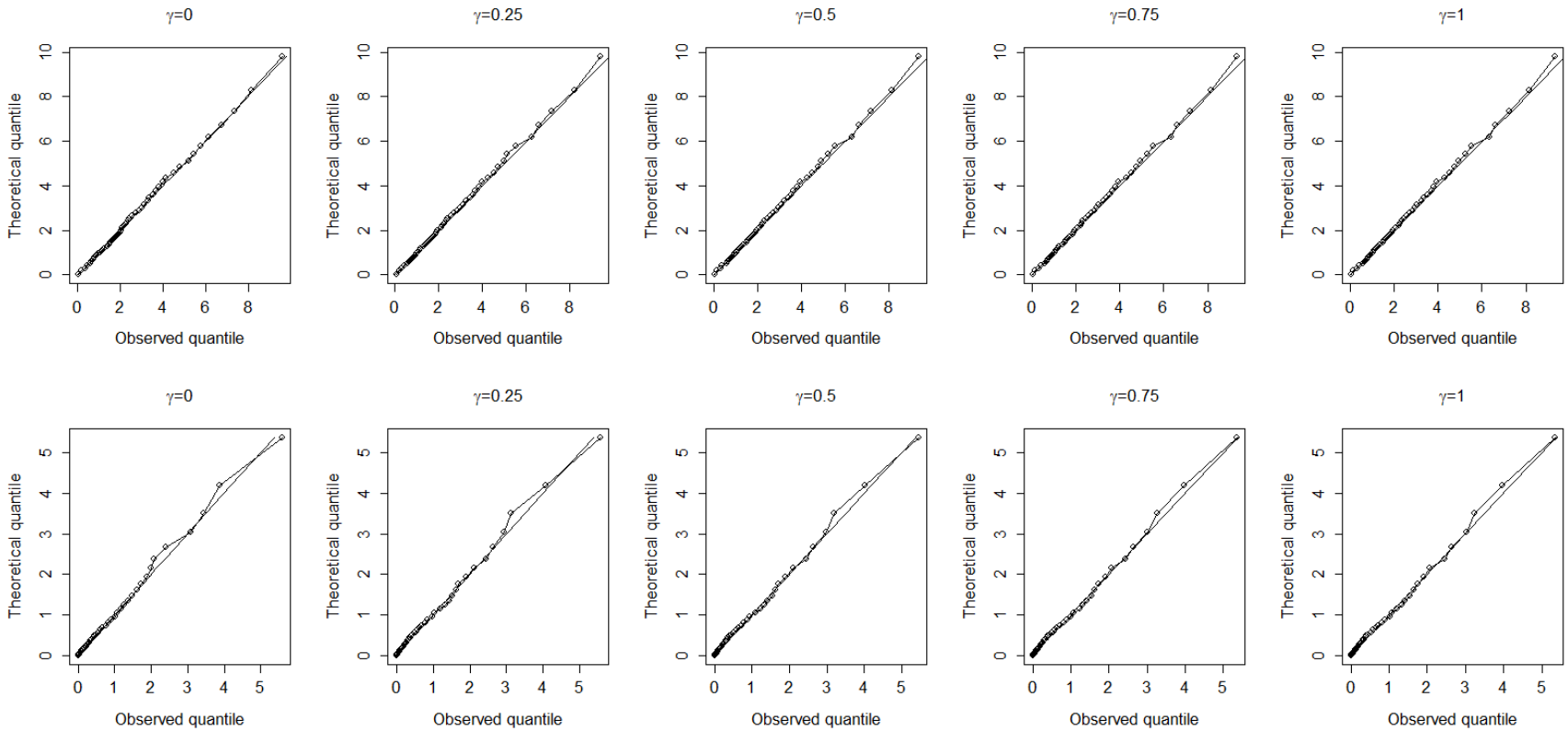


Figure 4.10: QQ-plots of the null distribution for HQIF ($H = H_{\mathrm{CL}}, \gamma = 0, 0.25, 0.5, 0.75, 1$) in 5-subregion network N3 (R^a_{CL}) with network size $m = 25$ and sample size $n = 500$. On the top panel: $\widehat{Q}_n(\widehat{\beta}|\gamma)$ relative to χ^2_3 and on the bottom panel $Q_n(a_0, \tilde{\beta}_B|\gamma) - Q_n(\widehat{\beta}_A, \widehat{\beta}_B|\gamma)$ relative to χ^2_1.

4.5 Data Example: infant's memory ERP study

We apply the proposed method to an infant's auditory recognition memory study conducted by the Center for Human Growth and Development. The electroencephalogram (EEG) data was recorded with a 64-channel HydroCel Geodesic Sensor Net for 161 infants of 2 months old, from which event-related potentials (ERP), a kind of neuroimaging data, are observed. Based on serum ferritin and ZPP levels in cord blood measured at birth, there were 52 infants labeled as being iron deficient (ID) while 109 infants are classified as iron sufficient (IS). Each infant hears both his/her mother's voice and stranger's voice in order to assess auditory recognition memory using EEG. The primary scientific objective of this study is to evaluate the effect of pre and/or postnatal environmental exposures (e.g. lead and pesticides) and iron deficiency (ID) on child neuro-developmental outcomes. After data pre-processing, data of 56 nodes are left for the analysis (see more details in Appendix E).

The outcome (y_{ij}) considered in this data analysis is a continuous variable of late slow wave (LSW) measured as a response to mother's voice stimulus, a kind of ERP reflective to memory updating. Nine covariates are included in the analysis. They are centered infant age (x_{i1}), centered lead (Pb) concentration in cord blood (x_{i2}), iron status (x_{i3}) as a binary measurement (with 1 for ID and 0 for IS), and six dummy variables for seven brain hemisphere regions, i.e. left frontal-central (x_{4j}), middle frontal-central (x_{5j}), right frontal-central (x_{6j}), left parietal-occipital (x_{7j}), middle parietal-occipital (x_{8j}), right parietal-occipital (x_{9j}) and other central (as the reference). More details about hemisphere regions and the average amplitude of LSW over the hemisphere are provided in Appendix. In this analysis, interaction effects between iron status and hemisphere regions (i.e. $x_{i3}x_{4j}$, $x_{i3}x_{5j}$, $x_{i3}x_{6j}$, $x_{i3}x_{7j}$, $x_{i3}x_{8j}$ and $x_{i3}x_{9j}$) are of key interest, as they enable us to assess whether iron status may alter the amplitude of LSW under mother's voice stimulus over the 7 regions. We

consider the marginal linear model of the following form:

$$
\begin{aligned}
E(y_{ij}|x_i) = \beta_0 &+ \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{4j} + \beta_5 x_{5j} + \beta_6 x_{6j} + \beta_7 x_{7j} \\
&+ \beta_8 x_{8j} + \beta_9 x_{9j} + \beta_{10} x_{i3} x_{4j} + \beta_{11} x_{i3} x_{5j} + \beta_{12} x_{i3} x_{6j} + \beta_{13} x_{i3} x_{7j} \\
&+ \beta_{14} x_{i3} x_{8j} + \beta_{15} x_{i3} x_{9j}, \ i = 1, \ldots, 161, \ j = 1, \ldots, 56.
\end{aligned} \tag{4.11}
$$

Table 4.6 reports the results of regression coefficient estimation, including point estimates, standard errors and sum of variance estimates obtained by several methods. They are GEE estimator with independent working network, HQIF estimator with fully data-driven structure HQIF($\gamma = 0$), HQIF estimators obtained under HQIF($H = H^*, \gamma = 1$) and those obtained from HQIF($H = H^*, \gamma = \widehat{\gamma}^*$).

In consultation with our collaborators, we consider two types of prior target H^* for the HQIF: one is a 7-block complete network $H_{7comp} = \text{block-diag}\{M_{\text{comp}}, \ldots, M_{\text{comp}}\}$ based on the 7-block hemisphere (see Fig. 4.1), and the other is a sparse network structure learned from the separate (or pilot) LSW data under stranger's voice stimulus using R package *space* (with a threshold 0.1), where the topology of H_{stranger} is displayed in Fig. 4.11.

As shown in Table 4.6, HQIF($H^* = H_{7\text{comp}}, \widehat{\gamma}^* = 0.875$) yields the smallest estimated total of variances $\text{tr}\{\widehat{\text{var}}(\widehat{\beta})\} = 1.263$ among five different HQIF estimators and the GEE estimator with independence network. The prior target $H_{7\text{comp}}$ is strongly favored with $\widehat{\gamma}^* = 0.875$ and thus highly informative to unveil the dependence of LSW outcomes among 56 nodes in comparison to the fully data-driven covariance matrix. The next top performer is HQIF($H^* = H_{\text{stranger}}, \widehat{\gamma}^* = 0.583$) with $\text{tr}\{\widehat{\text{var}}(\widehat{\beta})\} = 1.306$, suggesting that the prior target H_{stranger} is slightly more favorable than the data driven dependency structure with $\widehat{\gamma}^* = 0.583$. Although these two top methods provide similar parameter estimates, the former enables us to identify more significant group-region interaction effects than the latter. For example, interaction effect $\beta_{10} = 0.714$ is statistically significant, implying that the expected LSW amplitude is elevated by

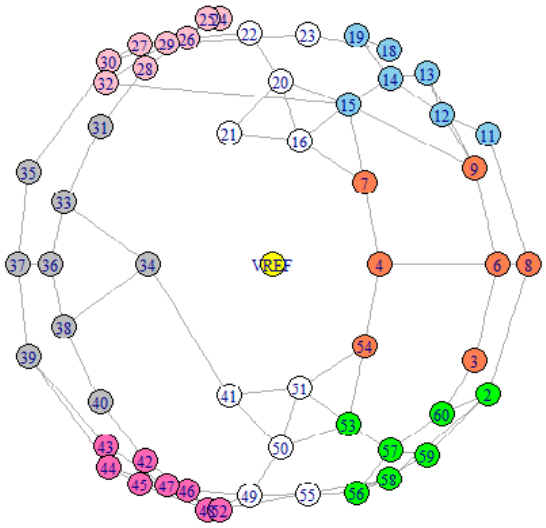


Figure 4.11: Sparse graphic representation of the learned network among 56 electrodes based on the LSW data under stranger's voice stimulus. Different colors of nodes represent 7 subregions.

0.714 units with the ID group over the IS group in the left frontal-central subregion. Likewise, significant interaction effect $\beta_{15} = -1.427$ suggests that the expected LSW amplitude is 1.427 units lower in the ID group than the IS group in the right parietal-occipital subregion of hemisphere. In summary, by allocating higher weights to more relevant network structure in the estimation and inference, the proposed HQIF($H = H^*, \gamma = \widehat{\gamma}^*$) method indicates promise to improve statistical power in the networked data analysis.

Table 4.6: The estimated regression parameters $\widehat{\beta}$ for the infant's memory ERP data to mother's voice stimulus(*: p-value<0.05). The estimated standard errors are reported inside the parentheses. The first four columns are HQIF estimators under two types of network structures suggested by our collaborators with different shrinkage coefficients. The other two columns are HQIF estimators with fully data-driven structure and GEE estimators with working independent network, respectively. "fc" denotes frontal-central and "po" denotes parietal-occipital. The last row lists the estimated sum of variance for $\widehat{\beta}$ (i.e. $\text{tr}\{\widehat{\text{var}}(\widehat{\beta})\}$). For HQIF method, $\text{tr}\{\widehat{\text{var}}(\widehat{\beta})\}$ is equivalent to $\widehat{\eta}(\gamma) = \text{tr}\left\{\widehat{J}^{-1}(\widehat{\beta}(\gamma)|\gamma, H^*)\right\}$ at $\gamma = 0, 1, \widehat{\gamma}^*$, where the $\widehat{\gamma}^*$ is determined based on the grid search method over a range from 0 to 1 with 25 equally spaced points.

	HQIF $H^* = H_{7\text{comp}}$		HQIF $H^* = H_{\text{stranger}}$		HQIF	GEE
parameter	$\widehat{\gamma}^*$=0.875	$\gamma = 1$	$\widehat{\gamma}^* = 0.583$	$\gamma = 1$	$\gamma = 0$	independence
age	-0.003 (0.002)	-0.001 (0.002)	-0.003 (0.001)	-0.006 (0.002)*	-0.003 (0.001)*	-0.001 (0.002)
Pb	-0.006 (0.003)	0 (0.004)	-0.005 (0.003)	0.003 (0.004)	-0.006 (0.003)*	0 (0.004)
group	0.158 (0.174)	0.502 (0.261)	0.158 (0.174)	0.247 (0.209)	0.176 (0.173)	0.587 (0.271)*
left fc	-0.803 (0.220)*	-0.794 (0.334)*	-0.854 (0.218)*	-0.483 (0.266)	-0.811 (0.220)*	-0.824 (0.335)*
middle fc	-0.360 (0.189)	-0.570 (0.273)*	-0.338 (0.183)	-0.535 (0.199)*	-0.363 (0.186)	-0.580 (0.275)*
right fc	-1.375 (0.218)*	-1.028 (0.341)*	-1.327 (0.215)*	-0.940 (0.228)*	-1.373 (0.218)*	-1.045 (0.343)*
left po	-0.167 (0.259)	0.494 (0.369)	-0.359 (0.251)	-0.764 (0.276)*	-0.200 (0.251)	0.466 (0.370)
middle po	-0.056 (0.281)	1.566 (0.367)*	-0.110 (0.280)	-0.567 (0.265)*	-0.071 (0.282)	1.566 (0.367)*
right po	0.573 (0.240)*	1.056 (0.392)*	0.603 (0.230)*	0.503 (0.252)*	0.559 (0.229)*	1.065 (0.392)*
group × left fc	0.714 (0.344)*	-0.672 (0.676)	0.571 (0.380)	-0.429 (0.474)	0.482 (0.374)	-1.143 (0.762)
group × middle fc	0.120 (0.312)	-0.641 (0.581)	0.092 (0.339)	-0.376 (0.380)	-0.081 (0.336)	-1.167 (0.703)
group × right fc	-0.462(0.379)	-0.604 (0.647)	-0.458 (0.388)	-0.456 (0.422)	-0.612 (0.390)	-1.101 (0.746)
group × left po	0.056 (0.392)	-0.086 (0.551)	0.167 (0.413)	0.192 (0.444)	0.246 (0.410)	0.207 (0.593)
group × middle po	-0.112(0.484)	-0.495 (0.665)	-0.080 (0.480)	0.313 (0.466)	-0.006 (0.491)	-0.277 (0.689)
group × right po	-1.427(0.374)*	-1.143 (0.677)	-1.417 (0.366)*	-1.165 (0.419)*	-1.337 (0.367)*	-0.959 (0.689)
$\text{tr}\{\widehat{\text{var}}(\widehat{\beta})\}$	1.263	3.238	1.306	1.568	1.314	3.763

4.6 Discussion

In this chapter, we have proposed a statistically efficient and computationally feasible approach to conducting the regression analysis of networked data. The proposed hybrid method constructs estimating functions based on the prior and data-driven network topology, and then combines them via the means of linear shrinkage. Compared with the existing GEE and QIF methods, our HQIF is more flexible and reliable to deal with complex dependence structures of networked data. Following the GMM theory, our method can naturally tune the shrinkage by minimizing the inverse of Godambe information, so it allocates higher weights to the more informative component of estimating function. Consequently, the HQIF improves the estimation efficiency over the existing GEE or QIF methods. The HQIF estimator has been also proven to be consistent and asymptotically Gaussian. In addition, HQIF-based test statistics follow chi-square distributions, which are not sensitive to the choice of tuning coefficient γ and the choice of the prior structure. This property allows one to enjoy the robustness of both the goodness-of-fit test and score-type test for nested models.

In practice although it is difficult to specify a very informative prior network topology, our simulation shows efficiency improvement as long as part of the prior structure captured is relevant. This is credited to the flexibility of our method that can accommodate not only a target structure but also the data-driven sample covariance. Note that, our methodology requires to estimate the common covariance V across the subjects. In practice, networked data may not be measured with the same size of vertices, and could be unbalanced due to data missingness or experimental constraints, such as bad channels in the EEG data. To improve the proposed method for unbalanced networked data, the sample covariance matrix could be possibly obtained by the method of *Qu et al.* (2010).

Finally, methods of sparse graph estimation are useful statistical tools to learn the target structure H from networked data. In practice, either training data or pilot

study data may not be always available. If the data is first analyzed to obtain the target H and then the same data is reanalyzed to obtain the results for regression model, we may face an over-fitting problem. In such situation, some adjustments may be needed to reach a proper inference. Nevertheless, the consistency of our HQIF estimation relies only on the unbiasedness assumption for the extended scores, which is not dependent on the choice of H and can be justified by the goodness-of-fit test provided in this chapter.

CHAPTER V

Summary and future work

This dissertation has focused on the development of statistical methodologies for the analysis of high-dimensional networked data. Due to complex dependence structures and large scales of such networked data, the need for innovative statistical models, analytic methods and algorithms has motivated me to the pursuit of three thesis research topics in this thesis.

In the first project, a sparse multivariate factor analysis regression model (smFARM) has been proposed in Chapter II to identify sparse association maps between gene expressions and biomarkers. The novelty of the proposed methodology lies on the utility of latent factors to segregate and quantify unobserved genetic variations from the measurement noise. The sparse learning method developed for the high-dimensional genetic data (large p small n) allows both the number of genes and the number of genetic variants to be high-dimensional, in which I proposed and implemented double sparsity penalties for both dimensions in order to yield desirable sparse association maps. As a result, my method can select a master regulator that relates to a group of genes and further identify non-zero associations of individual genes within a detected group including both cis-acting and trans-acting relationships. In addition, my method can also identify possible low-dimensional latent factors, and then evaluate and interpret the impact of latent factors on the association map by

gene-enrichment analysis.

In regard to the future work of Project I, many biomedical projects have collected time-course genetic variants and expression data, such as methylation data. Such longitudinal genetic/genomic data present new analytic challenges in modeling and data analysis, and furthermore, there is a need to generalize the association mapping models with random effects to evaluate the underlying genetic association maps for time-varying genetic data.

Chapter III discusses the reconstruction of gene regulatory networks (GRN) from expression data using sparse structural factor equation model (SFEM). The GRN may be formulated as a type of causal Gaussian network represented by directed acyclic graphs (DAGs). The general problem of estimation of directed graphs is computationally NP-hard and direction of interactions may not be distinguishable from observations. In the second project, I consider a special case of this problem, which relies on *a priori* knowledge of the ordering among network nodes. Such *a priori* knowledge of the ordering is usually provided by some existing annotation software such as Cytoscape, which is limited by context and tissue specificity. A future work of great importance is to generalize the assumption of the ordering, so that the ordering of variables is not prefixed. *Fu and Zhou* (2013), and *Aragam and Zhou* (2014) have done some work in this regard. They construct penalized estimation of DAGs from the structural equation models which circumvent the ordering problem by combining a method of enforcing acyclicity with a block coordinate descent algorithm. It will be interesting to see how the directions of causality among network nodes can be estimated under SFEM.

In the third project, a new regression analysis of networked data (RAND) approach has been proposed in Chapter IV to assess potential adverse effect of prenatal exposure to iron deficiency on auditory recognition memory of two-month old infants. RAND builds upon the quadratic inference functions (QIF) *Qu et al.* (2000); *Qu and*

Lindsay (2003), and further improves the estimating efficiency by incorporating the prior knowledge of network topology into constructing a flexible dependence model. The resulting hybrid quadratic inference functions (HQIF) take a similar form as the generalized method of moments *Hansen* (1982), hence preserves some desirable properties of GMM. For example, the HQIF estimator is consistent and asymptotically normally distributed. At the meanwhile, the HQIF objective function offers a means for hypothesis testing, such as a goodness-of-fit test and a score-type test for a nested model.

In terms of future work, two directions of research on EEG neuroimaging data are worth pursuing within the RAND framework. Note that the design of the cognitive study could be very complex, where ERPs could be recorded on the same subjects under different stimulus. To further adjust for the dependence structure among different stimulus, the first interesting problem is to extend RAND to analyze networked EEG data with repeated measurements. By decomposing the high-dimensional spatio-stimulus dependence structure into between-stimulus and within-stimulus components, an quasi-likelihood estimation approach can be developed, since the estimation efficiency is resulted from a construction of over-identified estimating equations.

In addition, the standard statistical analysis of ERPs is conducted on scalar measures such as mean or peak amplitude measured at prespecified location of EEG waves and does not take full advantage of the wealth of space and time information. Therefore, the second interesting problem is to generalize the RAND method by following the idea of the spatio-temporal correlation modeling. In this case, varying coefficient models (*Qu and Li*, 2006) can be useful, since this model can test whether coefficient functions are time varying or time invariant.

APPENDICES

APPENDIX A

M-fold Cross Validation

The M-fold cross-validation score CV_{ols} is given by:

$$\mathrm{CV}_{\mathrm{ols}}(\lambda_1, \lambda_2; K = K_0) = \sum_{i=1}^{M} \mathrm{trace}\Big\{ \big(Y^{(i)} - X^{(i)}\widetilde{\Theta}^{(i)^T}\big)\big(\widehat{\Sigma}_{\mathrm{ols}}^{(i)}\big)^{-1}\big(Y^{(i)} - X^{(i)}\widetilde{\Theta}^{(i)^T}\big)^T \Big\}, \tag{A.1}$$

where $\widetilde{\Theta}^{(i)} = \widetilde{\Theta}^{(i)}(\lambda_1, \lambda_2)$ is the OLS estimates based on the i-th training data $S^{-(i)} = (Y^{-(i)}, X^{-(i)})$, i.e. the subset of the data with the i-th sample deleted. And $\widetilde{\Theta}^{(i)}(\lambda_1, \lambda_2) = \{\tilde{\theta}_{qp}^{(i)}\}$ is derived by taking the following steps:

- Given the shrunken estimates $\hat{\Theta}^{(i)}(\lambda_1, \lambda_2) = \{\hat{\theta}_{qp}^{(i)}\}$, for each response q, define a $Y_q^{-(i)}$-oriented active predictor set $\mathscr{H}_q = \{p : \hat{\theta}_{qp}^{(i)} \neq 0, 1 \leq p \leq P\}$;
- Set $\tilde{\theta}_{qp}^{(i)} = 0$, if $p \notin \mathscr{H}_q$; otherwise, $\{\tilde{\theta}_{qp}^{(i)}, p \in \mathscr{H}_q\}$ contains the OLS estimates obtained by regressing $Y_q^{-(i)}$ on $\{X_p^{-(i)}, p \in \mathscr{H}_q\}$.

In (A.1), $\widehat{\Sigma}_{\mathrm{ols}}^{(i)}$ is calculated from the factor analysis model using $S^{-(i)}$ given $\widetilde{\Theta}^{(i)}(\lambda_1, \lambda_2)$. Hereafter, the optimal tuning parameters $(\lambda_1^*, \lambda_2^*)$ at $K = K_0$ are determined by minimizing $\mathrm{CV}_{\mathrm{ols}}(\lambda_1, \lambda_2; K = K_0)$. And the optimal number of latent factors K_{CV} are chosen by minimizing $\mathrm{CV}_{\mathrm{ols}}(\lambda_1^*, \lambda_2^*; K = K_0)$ among $K_0 = 0$ to a given large number.

APPENDIX B

Proof of Proposition II.1

For notational convenience in presenting the algorithm, we first set $\tilde{E} = ZB^T + E$ and reformulate model $Y = X\Theta^T + ZB^T + E$ as the following format

$$\underset{NQ\times 1}{\overrightarrow{\mathbb{Y}}} = \underset{NQ\times QPQP\times 1}{\mathbb{X}}\ \underset{}{\overrightarrow{\Gamma}} + \underset{NQ\times 1}{\overrightarrow{\mathbb{E}}}, \tag{B.1}$$

where $\overrightarrow{\mathbb{Y}} \triangleq \text{Vec}(Y^T)$, $\mathbb{X} \triangleq X \otimes I_Q = (X_1 \otimes I_Q, \ldots, X_P \otimes I_Q)$, $\overrightarrow{\mathbb{E}} \triangleq \text{Vec}(\tilde{E}^T) = (Z \otimes I_Q)\text{Vec}(B) + \text{Vec}(E^T)$. Note that, $\mathbb{W} \triangleq Var(\overrightarrow{\mathbb{Y}}) = I_N \otimes (BB^T + \Psi) = I_N \otimes \Sigma$, with $\mathbb{W}^{-1} = I_N \otimes \Sigma^{-1}$, $\overrightarrow{\Gamma} \triangleq \text{Vec}(\Theta) = (\theta_1^T, \ldots, \theta_P^T)^T$ with $\theta_p = (\theta_{1p}, \ldots, \theta_{Qp})^T$, and $\overrightarrow{\mathbb{C}} = \text{Vec}(C) = (C_1^T, \ldots, C_P^T)^T$ with $C_p = (C_{1p}, \ldots, C_{Qp})^T$, our corresponding joint least squares loss function is given by

$$\begin{aligned} L(\overrightarrow{\Gamma}; \mathbb{W}, \lambda_1, \lambda_2) = \frac{1}{2N}\Big[\overrightarrow{\mathbb{Y}} - \sum_{p=1}^{P}(X_p \otimes I_Q)\theta_p\Big]^T \mathbb{W}^{-1}\Big[\overrightarrow{\mathbb{Y}} - \sum_{p=1}^{P}(X_p \otimes I_Q)\theta_p\Big] \\ + \lambda_1 \sum_{p=1}^{P} \|C_p \odot \theta_p\|_1 + \lambda_2 \sum_{p=1}^{P} \|C_p \odot \theta_p\|_2, \end{aligned}$$

where $\odot$ represents the element-wise operator Hadamard product. In order to show Proposition II.1, given $[\Theta(\cdot, p_0)]$, we first split θ_{p_0} into two parts $\theta_{p_0}^{\mathscr{A}}$ and $\theta_{p_0}^{\mathscr{B}}$. Next, we will iterate between $\theta_{p_0}^{\mathscr{A}}$ and $\theta_{p_0}^{\mathscr{B}}$ to update θ_{p_0}.

Given the current estimates of Θ, $\hat{\theta}^{\mathscr{A}}_{p_0}$ is the solution of the following constrained optimization problem:

$$\begin{aligned}\arg\min_{\beta}\frac{1}{2N}\Big[\mathrm{Vec}(\widetilde{Y}^T_{\mathscr{A}_{p_0}})-(X_{p_0}\otimes I_Q)\beta\Big]^T\mathbb{W}^{-1}\Big[\mathrm{Vec}(\widetilde{Y}^T_{\mathscr{A}_{p_0}})-(X_{p_0}\otimes I_Q)\beta\Big],\\ \text{s. t. } H_{\mathscr{A}_{p_0}}\beta=0.\end{aligned} \tag{B.2}$$

where $\widetilde{Y}^T_{\mathscr{A}_{p_0}}=Y^T-[\Theta(\cdot,p_0)]X^T-\theta^{\mathscr{B}}_{p_0}X^T_{p_0}$, and $H_{\mathscr{A}_{p_0}}$ is a $\|C_{p_0}\|_0\times Q$ matrix of full rank with elements 0 or 1, satisfying $H_{\mathscr{A}_{p_0}}\theta^{\mathscr{A}}_{p_0}=0$, which sets $\theta^{\mathscr{A}}_{qp_0}=0$ for $q\notin\mathscr{A}_{p_0}$. Employing the method of Lagrange multipliers, we get the following constraint estimator of $\theta^{\mathscr{A}}_{p_0}$

$$\hat{\theta}^{\mathscr{A}}_{p_0}=\Big\{I_Q-\Sigma H^T_{\mathscr{A}_{p_0}}\Big(H_{\mathscr{A}_{p_0}}\Sigma H^T_{\mathscr{A}_{p_0}}\Big)^{-1}H_{\mathscr{A}_{p_0}}\Big\}\widetilde{Y}^T_{\mathscr{A}_{p_0}}X_{p_0}/\big\|X_{p_0}\big\|_2^2, \tag{B.3}$$

On the other hand, $\hat{\theta}^{\mathscr{B}}_{p_0}$ is the solution of the constrained optimization problem:

$$\begin{aligned}\arg\min_{\beta}\frac{1}{2N}\Big[\mathrm{Vec}(\widetilde{Y}^T_{\mathscr{B}_{p_0}})-(X_{p_0}\otimes I_Q)\beta\Big]^T\mathbb{W}^{-1}\Big[\mathrm{Vec}(\widetilde{Y}^T_{\mathscr{B}_{p_0}})-(X_{p_0}\otimes I_Q)\beta\Big]\\ +\lambda_1\|C_{p_0}\odot\beta\|_1+\lambda_2\|C_{p_0}\odot\beta\|_2,\\ \text{s. t. } H_{\mathscr{B}_{p_0}}\beta=0.\end{aligned} \tag{B.4}$$

where $\widetilde{Y}^T_{\mathscr{B}_{p_0}}=Y^T-[\Theta(\cdot,p_0)]X^T-\theta^{\mathscr{A}}_{p_0}X^T_{p_0}$, and $H_{\mathscr{B}_{p_0}}$ is a $(Q-\|C_{p_0}\|_0)\times Q$ matrix with 0 and 1 elements, which set $\theta^{\mathscr{B}}_{qp_0}=0$, if $q\notin\mathscr{B}_{p_0}$. The subgradient equation of $\theta^{\mathscr{B}}_{p_0}$ is given by

$$\theta^{\mathscr{B}}_{p_0}=\Big\{I_Q-\Sigma H^T_{\mathscr{B}_{p_0}}\Big(H_{\mathscr{B}_{p_0}}\Sigma H^T_{\mathscr{B}_{p_0}}\Big)^{-1}H_{\mathscr{B}_{p_0}}\Big\}\Big[\widetilde{Y}^T_{\mathscr{B}_{p_0}}X_{p_0}-N\Sigma(\lambda_1 s_{p_0}+\lambda_2 t_{p_0})\Big]/\|X_{p_0}\|_2^2, \tag{B.5}$$

where $s_{p_0} = (s_{1p_0}, \ldots, s_{Qp_0})^T$, with

$$s_{qp_0} = \begin{cases} sgn(\theta_{qp_0}), \text{ if } q \in \mathscr{B}_{p_0} \text{ and } \theta_{qp_0} \neq 0, \\ \in [-1, 1], \text{ if } q \in \mathscr{B}_{p_0} \text{ and } \theta_{qp_0} = 0, \\ 0, \text{ if } q \notin \mathscr{B}_{p_0}, \end{cases} \tag{B.6}$$

and $t_{p_0} = (t_{1p_0}, \ldots, t_{Qp_0})^T$, with

$$t_{qp_0} = \begin{cases} \theta_{qp_0} / \|\theta_{p_0}^{\mathscr{B}}\|_2, \text{ if } q \in \mathscr{B}_{p_0} \text{ and } \|\theta_{p_0}^{\mathscr{B}}\|_2 \neq 0, \\ 0, \quad \text{if } q \notin \mathscr{B}_{p_0}, \end{cases} \tag{B.7}$$

and $\|t_{p_0}\|_2 \leq 1$, if $\|\theta_{p_0}^{\mathscr{B}}\|_2 = 0$. If $\|\theta_{p_0}^{\mathscr{B}}\|_2 = 0$, (B.5) is equivalent to $\frac{1}{N}\widetilde{Y}_{\mathscr{B}_{p_0}}^T X_{p_0} - \lambda_1 \Sigma s_{p_0} = \lambda_2 \Sigma t_{p_0}$. We can determine it by minimizing $J(s_{p_0}) = \|\frac{1}{N}\Sigma^{-1}\widetilde{Y}_{\mathscr{B}_{p_0}}^T X_{p_0} - \lambda_1 s_{p_0}\|_2$ with respect to s_{p_0} and check if $J(\hat{s}_{p_0}) \leq \lambda_2$, with

$$\hat{s}_{qp_0} = \begin{cases} sgn\Big(\frac{1}{\lambda_1 N} X_{p_0}^T \widetilde{Y}_{\mathscr{B}_{p_0}} \Sigma_q^{-1}\Big) \min\Big(|\frac{1}{\lambda_1 N} X_{p_0}^T \widetilde{Y}_{\mathscr{B}_{p_0}} \Sigma_q^{-1}|, 1\Big), \text{ if } q \in \mathscr{B}_{p_0}, \\ 0, \text{ if } q \notin \mathscr{B}_{p_0}. \end{cases}$$

If $J(\hat{s}_{p_0}) \leq \lambda_2$, we have $\hat{\theta}_{p_0}^{\mathscr{B}} = 0_{Q \times 1}$; otherwise, we will use a coordinate-wise algorithm to update θ_{qp_0} if $q \in \mathscr{B}_{p_0}$. Given $q_0 \in \mathscr{B}_{p_0}$ and a $Q \times P$ matrix $[\Theta(q_0, p_0)]$, the subgradient equation of $\theta_{q_0 p_0}$ is

$$\begin{aligned} -\frac{1}{N} X_{p_0}^T \big(Y - X[\Theta(q_0, p_0)]^T\big) \Sigma_{q_0}^{-1} + \frac{\|X_{p_0}\|_2^2}{N} \Sigma_{q_0 q_0}^{-1} \theta_{q_0 p_0} \\ + \lambda_1 s_{q_0 p_0} + \lambda_2 t_{q_0 p_0} = 0, \end{aligned} \tag{B.8}$$

where

$$s_{q_0 p_0} = \begin{cases} sgn(\theta_{q_0 p_0}), \text{ if } \theta_{q_0 p_0} \neq 0, \\ \in [-1, 1], \text{ if } \theta_{q_0 p_0} = 0, \end{cases} \tag{B.9}$$

$$t_{q_0p_0} = \begin{cases} \dfrac{\theta_{q_0p_0}}{\sqrt{\theta^2_{q_0p_0} + \left\|[\theta^{\mathscr{B}}_{p_0}(q_0)]\right\|_2^2}}, \text{ if } \left\|[\theta^{\mathscr{B}}_{p_0}(q_0)]\right\|_2 \neq 0, \\ sgn(\theta_{q_0p_0}), \text{ if } \left\|[\theta^{\mathscr{B}}_{p_0}(q_0)]\right\|_2 = 0 \text{ and } \theta_{q_0p_0} \neq 0, \\ \in [-1,1], \text{ if } \left\|[\theta^{\mathscr{B}}_{p_0}(q_0)]\right\|_2 = 0 \text{ and } \theta_{q_0p_0} = 0\ . \end{cases} \tag{B.10}$$

Therefore, from (B.8) we have

$$\hat{\theta}_{q_0p_0} = \begin{cases} \dfrac{NS\Big(\frac{1}{N}X^T_{p_0}\big(Y - X[\Theta(q_0,p_0)]^T\big)\Sigma^{-1}_{q_0},\ \lambda_1+\lambda_2\Big)}{\Sigma^{-1}_{q_0q_0}\left\|X_{p_0}\right\|_2^2}, \text{ if } \left\|[\theta^{\mathscr{B}}_{p_0}(q_0)]\right\|_2 = 0, \\ 0, \text{ if } \left|\dfrac{1}{N}X^T_{p_0}\big(Y - X[\Theta(q_0,p_0)]^T\big)\Sigma^{-1}_{q_0}\right| \leq \lambda_1, \text{ and } \left\|[\theta^{\mathscr{B}}_{p_0}(q_0)]\right\|_2 \neq 0. \end{cases} \tag{B.11}$$

Otherwise, $\hat{\theta}_{q_0p_0}$ is the solution of the following nonlinear equation:

$$\begin{aligned} &-\frac{1}{N}X^T_{p_0}\big(Y - X[\Theta(q_0,p_0)]^T\big)\Sigma^{-1}_{q_0} + \frac{\|X_{p_0}\|_2^2}{N}\Sigma^{-1}_{q_0q_0}\theta_{q_0p_0} \\ &+ \lambda_1 sgn(\theta_{q_0p_0}) + \lambda_2\frac{\theta_{q_0p_0}}{\sqrt{\theta^2_{q_0p_0} + \left\|[\theta^{\mathscr{B}}_{p_0}(q_0)]\right\|_2^2}} = 0. \end{aligned} \tag{B.12}$$

Since equation (B.12) is nonlinear with respect to $\theta_{q_0p_0}$, an approximate solution to (B.12) can be given by approximating the group lasso penalty $\sqrt{\theta^2_{q_0p_0} + \left\|[\theta^{\mathscr{B}}_{p_0}(q_0)]\right\|_2^2}$ with a ridge penalty $\frac{\theta^2_{q_0p_0} + \left\|[\theta^{\mathscr{B}}_{p_0}(q_0)]\right\|_2^2}{\left\|\theta^{\mathscr{B}}_{p_0}\right\|_2}$. This results in the following solution:

$$\hat{\theta}_{q_0p_0} = \frac{NS\Big(\frac{1}{N}X^T_{p_0}\big(Y - X[\Theta(q_0,p_0)]^T\big)\Sigma^{-1}_{q_0},\ \lambda_1\Big)}{\Sigma^{-1}_{q_0q_0}\left\|X_{p_0}\right\|_2^2 + 2N\lambda_2\left\|\theta^{\mathscr{B}}_{p_0}\right\|_2^{-1}}, \tag{B.13}$$

where $[\Theta(q_0,p_0)]$ and $\theta^{\mathscr{B}}_{p_0}$ are obtained from the current estimates of Θ.

APPENDIX C

Additional details concerning analysis of gene set enrichment

GSEA (`http://www.broadinstitute.org/gsea/msigdb/help_annotations.jsp#overlap`) and Gather (`http://gather.genome.duke.edu/`) are two web-based softwares. To assess enrichment for specific lists of probe sets, we simply enter the list into the web based interface and all associations are tested and reported.

Here we fit smFARM with and without incorporating the CNAI information and compute overlaps of a set of factor-driven genes with CP, GO and KEGG pathways. Here factor-driven genes are genes satisfying factor loadings $|\hat{B}_{q,k}| > 0.1$ after varimax rotation. Results of enrichment are summarized in Table C.1.

From Table C.1, we find that adjusting for the variations of CNAIs allows us to detect more enriched gene sets as well as to obtain smaller p-values for pathways. This again indicates that simultaneously considering gene set associations from latent factors and the shared genetic effects from CNAIs can improve the efficiency in the association analysis.

Table C.1: Summary of enrichment tests for 654 breast cancer related genes with or without CNAIs effects (Top overlaps $p < 1.00e^{-5}$).

			Adjusting for CNAIs effect		Ignoring CNAIs effect	
GO ID	GO Term	No. Genes in Gene Set	No. Genes in Overlap	p-value	No. Genes in Overlap	p-value
0007049	CELL CYCLE	315	40	$5.73e^{-10}$	39	$2.11e^{-09}$
0000278	MITOTIC CELL CYCLE	153	26	$1.43e^{-09}$	25	$7.14e^{-09}$
0022402	CELL CYCLE PROCESS	193	29	$3.13e^{-09}$	28	$1.36e^{-08}$
0022403	CELL CYCLE PHASE	170	25	$6.02e^{-07}$	24	$2.52e^{-07}$
0000279	M PHASE	114	18	$1.52e^{-06}$	18	$1.56e^{-06}$
0000087	M PHASE OF MITOTIC CELL CYCLE	85	15	$2.73e^{-06}$	15	$2.80e^{-06}$
0005694	CHROMOSOME	124	18	$5.22e^{-06}$	—	$> 1.00e^{-05}$
0007067	MITOSIS	82	14	$8.74e^{-06}$	14	$8.95e^{-06}$
KEGG ID	KEGG pathway					
map00480	GLUTATHIONE METABOLISM	50	11	$6.50e^{-06}$	11	$6.63e^{-06}$
CP ID	Canonical pathways					
—	REACTOME CELL CYCLE MITOTIC	325	34	$1.29e^{-06}$	34	$3.73e^{-06}$
—	PID FOXM1PATHWAY	40	10	$5.03e^{-06}$	—	$> 1.00e^{-05}$

Note: Top overlaps with cutoff $p < 1.00e^{-5}$ are chosen because the results from another web-based statistical tool Gather [see Chang et al. (2006), `http://gather.genome.duke.edu/`] are highly matched these top gene sets from GSEA when $p < 1.00e^{-5}$.

APPENDIX D

Proof of Lemma IV.1

Given a target structure H, and under some regularity conditions stated in *Harris et al.* (1999)'s Chapter 1, for a given γ, $\widehat{\beta}(\gamma)$ obtained by minimising the HQIF (4.8) given in Chapter 4 is consistent and asymptotically normal (see Section 4.3.2). In addition, since the weighting covariance matrix $\bar{\Gamma}_n(\widehat{\beta}|\gamma) \xrightarrow{p} \Gamma(\beta_0|\gamma)$ and $\dot{\bar{g}}_n(\widehat{\beta}|\gamma) \xrightarrow{p} G(\beta_0|\gamma)$, the inverse of the Godambe information matrix $J^{-1}(\beta_0|\gamma)$ of g_i may be consistently estimated by $\widehat{J}^{-1}(\widehat{\beta}(\gamma)|\gamma) = \big\{\dot{\bar{g}}_n^T(\widehat{\beta}(\gamma)|\gamma)\bar{\Gamma}_n^{-1}(\widehat{\beta}(\gamma)|\gamma)\dot{\bar{g}}_n(\widehat{\beta}(\gamma)|\gamma)\big\}^{-1}$, so is its trace, i.e. $\operatorname{tr}\{\widehat{J}^{-1}(\widehat{\beta}(\gamma)|\gamma)\} \xrightarrow{p} \operatorname{tr}\{J^{-1}(\beta_0|\gamma)\}$.

Let $\widehat{\eta}(\gamma) = \operatorname{tr}\{\widehat{J}^{-1}(\widehat{\beta}(\gamma)|\gamma)\}$, and let $\eta_0(\gamma)$ be $\operatorname{tr}\{J^{-1}(\beta_0|\gamma)\}$. It follows that $\widehat{\eta}(\gamma) - \eta_0(\gamma) \xrightarrow{p} 0$ pointwise of γ on the compact set $[0, 1]$. To show that $\widehat{\eta}(\gamma)$ is stochastically equicontinuous, we check a stochastic Lipschitz-type condition on $\widehat{\eta}(\gamma)$, $E\{sup_{\gamma\in[0,1]}|\partial\widehat{\eta}(\gamma)/\partial\gamma|\} < \infty$. With the application of Lemma 1 in *Wang et al.* (1986), we have

$$\begin{aligned}\frac{\partial\widehat{\eta}(\gamma)}{\partial\gamma} &= -\operatorname{tr}\{\widehat{J}^{-1}(\gamma)\widehat{W}(\gamma)\widehat{J}^{-1}(\gamma)\} = -\operatorname{tr}\{\widehat{W}(\gamma)\widehat{J}^{-2}(\gamma)\} \\ &\leq \rho\big(\widehat{W}(\gamma)\big)\operatorname{tr}\{\widehat{J}^{-2}(\gamma)\} \leq \|\widehat{W}(\gamma)\|_{m_\infty}\operatorname{tr}\{\widehat{J}^{-2}(\gamma)\},\end{aligned} \tag{D.1}$$

where $\widehat{W}(\gamma)$ is given in (D.2) below and $\rho\big(\widehat{W}(\gamma)\big)$ is the spectral radius of a $p \times p$ real symmetric matrix $\widehat{W}(\gamma)$ and $\|\cdot\|_{m_\infty}$ is a matrix norm defined as $\|\widehat{W}\|_{m_\infty} =$

$p \cdot \max_{i,j} |\widehat{W}_{ij}|$. Note that,

$$\widehat{W}(\gamma) \stackrel{\text{def}}{=} \frac{\partial \widehat{J}(\widehat{\beta}(\gamma)|\gamma)}{\partial \gamma} = \frac{\partial \dot{\bar{g}}_n^{*T}}{\partial \gamma} \bar{\Gamma}_n^{-1} \dot{\bar{g}}_n + \dot{\bar{g}}_n^{*T} \bar{\Gamma}_n^{-1} \frac{\partial \dot{\bar{g}}_n}{\partial \gamma} - \dot{\bar{g}}_n^{*T} \bar{\Gamma}_n^{-1} \frac{\partial \bar{\Gamma}_n}{\partial \gamma} \bar{\Gamma}_n^{-1} \dot{\bar{g}}_n, \tag{D.2}$$

with $\frac{\partial \dot{\bar{g}}_n}{\partial \gamma} = \dot{\bar{f}}_n - \dot{\bar{h}}_n$ and $\dot{\bar{g}}_n = \gamma \dot{\bar{f}}_n + (1-\gamma)\dot{\bar{h}}_n$. For sufficiently large n, $\dot{\bar{g}}_n$ is continuous on $\gamma \in [0,1]$ and thus bounded. Since $\bar{\Gamma}_n$ is also bounded and positive definite on $\gamma \in [0,1]$, so does the $\bar{\Gamma}_n^{-1}$. Besides, the expression of $\bar{\Gamma}_n(\gamma) = n^{-1} \sum_{i=1}^{n} \big\{\gamma f_i + (1-\gamma)h_i\big\}\big\{\gamma f_i + (1-\gamma)h_i\big\}^T$ implies that $\partial \bar{\Gamma}_n(\gamma)/\partial\gamma$ is continuous on [0, 1], so $\partial \bar{\Gamma}_n(\gamma)/\partial\gamma$ is also element-wise bounded. Hence, each term in (D.2) is elementwise bounded on $\gamma \in [0,1]$ and we have $\|\widehat{W}(\gamma)\|_{m_\infty} < \infty$. On the other hand, the regularity conditions ensure that $\text{tr}\{\widehat{J}^{-2}(\gamma)\} < \infty$. Thus, $\frac{\partial \widehat{\eta}(\gamma)}{\partial \gamma}$ can be uniformly bounded, and the Lipschitz-type condition is satisfied. Then we have the uniformity of convergence *Newey* (1991), $sup_{\gamma\in[0,1]}|\widehat{\eta}(\gamma) - \eta_0(\gamma)| \xrightarrow{p} 0$. Finally, since $S_0 = \{\gamma : \max_{\gamma\in[0,1]} \eta_0(\gamma)\}$ and $S = \{\gamma : \max_{\gamma\in[0,1]} \widehat{\eta}(\gamma)\}$ with $|S_0| = |S| < \infty$, we have $\widehat{\gamma}^* \xrightarrow{p} \gamma_0^*$, as $n \to \infty$, where $\gamma_0^* = sup\ \{S_0\}$ and $\widehat{\gamma}^* = sup\ \{S\}$.

APPENDIX E

Additional background about the infant memory ERP study

The electroencephalogram (EEG) was recorded by a 64-electrode HydroCel Geodesic Sensor Net (Electrical Geodesics Inc., Eugene, OR). Due to data quality, the following electrodes are excluded in the data analysis: (62, 63), (1, 17, 61, 64), (5, 10). The remaining 56 electrodes are used in the analysis. Our collaborators outlined six regions of interests for waveform analysis, including left frontal-central (11, 12, 13, 14, 15, 18, 19), middle frontal-central (3, 4, 6, 7, 8, 9, 54), right frontal-central (2, 53, 56, 57, 58, 59, 60), left parietal-occipital (24, 25, 26, 27, 28, 29, 30, 32), middle parietal-occipital (31, 33, 34, 35, 36, 37, 38, 39, 40), right parietal-occipital (42, 43, 44, 45 46, 47, 48, 52), and the seventh region containing the other central nodes (16, 20, 21, 22, 23, 41, 49, 50, 51, 55). See Figure 4.1 for their positions in Chapter 4.

Fig. E.1 displays the average amplitude of LSW under mother's voice stimulus for IS group (the left panel) or ID group (the right panel) over the hemisphere. From this figure we see that the average amplitude of LSW in either IS group or ID group is negative in both frontal and central regions (left frontal-central, middle frontal-central and right frontal-central), but positive in parietal-occipital scalp regions (left parietal-occipital, middle parietal-occipital and right parietal-occipital). The objective of

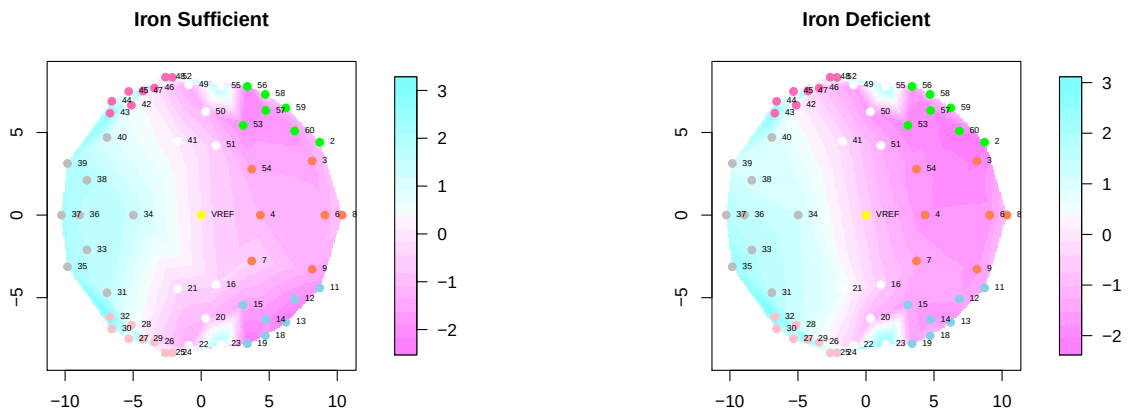


Figure E.1: The average amplitude of LSW under mother's voice stimulus for each iron group.

the data analysis is to assess interaction effects between iron status and hemisphere regions, adjusting for confounding factors.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ahmad, A., Z. W. Wang, D. J. Kong, S. Ali, Y. W. Li, S. Banerjee, R. Ali, and F. H. Sarkar (2010), Foxm1 down-regulation leads to inhibition of proliferation, migration and invasion of breast cancer cells through the modulation of extra-cellular matrix degrading factors, *Breast Cancer Research and Treatment*, *122*(2), 337–346.

Ahn, S. C., and A. R. Horenstein (2013), Eigenvalue ratio test for the number of factors, *Econometrica*, *81*(3), 1203–1227.

Allison, D. B., X. Cui, G. P. Page, and M. Sabripour (2006), Microarray data analysis: from disarray to consolidation and consensus, *Nature Reviews Genetics*, *7*(1), 55–65.

Alter, O., P. O. Brown, and D. Botstein (2000), Singular value decomposition for genome-wide expression data processing and modeling, *Proceedings of the National Academy of Sciences of the United States of America*, *97*(18), 10,101–10,106.

Altomare, D., G. Consonni, and L. La Rocca (2013), Objective bayesian search of gaussian directed acyclic graphical models for ordered variables with nonlocal priors, *Biometrics*, *69*(2), 478–487.

Anandkumar, A., D. Hsu, A. Javanmard, and S. Kakade (2013), Learning linear bayesian networks with latent variables, in *Proceedings of The 30th International Conference on Machine Learning*, pp. 249–257.

Anderson, T. W., and H. Rubin (1956), Statistical inference in factor analysis, *In Proceedings of the third Berkeley symposium on mathematical statistics and probability*, *5*, 111–150.

Aragam, B., and Q. Zhou (2014), Concave penalized estimation of sparse bayesian networks, *arXiv preprint arXiv:1401.0852*.

Bai, J., and S. Ng (2002), Determining the number of factors in approximate factor models, *Econometrica*, *70*(1), 191–221.

Bansal, V., O. Libiger, A. Torkamani, and N. J. Schork (2010), Statistical analysis strategies for association studies involving rare variants, *Nature Reviews Genetics*, *11*(11), 773–785.

Basso, K., A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano (2005), Reverse engineering of regulatory networks in human b cells, *Nature genetics*, *37*(4), 382–390.

Bedrick, E. J., and C. L. Tsai (1994), Model selection for multivariate regression in small samples, *Biometrics*, *50*(1), 226–231.

Bhatia, R. (1997), *Matrix Analysis*, Springer New York.

Bickel, P. J., and E. Levina (2008), Covariance regularization by thresholding, *Annals of Statistics*, *36*(6), 2577–2604.

Blum, Y., G. Le Mignon, S. Lagarrigue, and D. Causeur (2010), A factor model to analyze heterogeneity in gene expression, *Bmc Bioinformatics*, *11*.

Breiman, L. (1995), Better subset regression using the nonnegative garrote, *Technometrics*, *37*(4), 373–384.

Brem, R. B., and L. Kruglyak (2005), The landscape of genetic complexity across 5,700 gene expression traits in yeast, *Proceedings of the National Academy of Sciences of the United States of America*, *102*(5), 1572–1577.

Butte, A. J., P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane (2000), Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks, *Proceedings of the National Academy of Sciences*, *97*(22), 12,182–12,186.

Cai, T. T., H. Z. Li, W. D. Liu, and J. C. Xie (2013), Covariate-adjusted precision matrix estimation with an application in genetical genomics, *Biometrika*, *100*(1), 139–156.

Caner, M., and X. Han (2013), Selecting the correct number of factors in approximate factor models: The large panel case with group bridge estimators, *Journal of Business and Economic Statistics (conditionally accepted)*.

Carvalho, C. M., J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West (2008), High-dimensional sparse factor modeling: applications in gene expression genomics, *Journal of the American Statistical Association*, *103*(484).

Chang, J. T., and J. R. Nevins (2006), Gather: a systems approach to interpreting genomic signatures, *Bioinformatics*, *22*(23), 2926–2933.

Cheng, J., R. Greiner, J. Kelly, D. Bell, and W. R. Liu (2002), Learning bayesian networks from data: An information-theory based approach, *Artificial Intelligence*, *137*(1-2), 43–90, cheng, J Greiner, R Kelly, J Bell, D Liu, WR.

Dobra, A., C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West (2004), Sparse graphical models for exploring gene expression data, *Journal of Multivariate Analysis*, *90*(1), 196–212.

Drton, M., R. Foygel, and S. Sullivant (2011), Global identifiability of linear structural equation models, *The Annals of Statistics*, *39*(2), 865–886.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004), Least angle regression, *Annals of Statistics*, *32*(2), 407–451.

Ellis, B., and W. H. Wong (2008), Learning causal bayesian network structures from experimental data, *Journal of the American Statistical Association*, *103*(482), 778–789, ellis, Byron Wong, Wing Hung.

Fan, J., Y. Feng, and Y. Wu (2009), Network exploration via the adaptive lasso and scad penalties, *The annals of applied statistics*, *3*(2), 521.

Fan, J. Q., and R. Z. Li (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fields, E. C., and G. R. Kuperberg (2012), It's all about you: An erp study of emotion and self-relevance in discourse, *Neuroimage*, *62*(1), 562–574.

Fretham, S. J., E. S. Carlson, and M. K. Georgieff (2011), The role of iron in learning and memory, *Advances in Nutrition: An International Review Journal*, *2*(2), 112–121.

Friedman, J., T. Hastie, H. Hfling, and R. Tibshirani (2007), Pathwise coordinate optimization, *The Annals of Applied Statistics*, *1*(2), 302–332.

Friedman, J., T. Hastie, and R. Tibshirani (2008), Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, *9*(3), 432–441.

Friedman, J., T. Hastie, and R. Tibshirani (2010a), Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, *33*(1), 1–22.

Friedman, J., T. Hastie, and R. Tibshirani (2010b), A note on the group lasso and a sparse group lasso, *arXiv preprint arXiv:1001.0736*.

Friedman, N., and D. Koller (2003), Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks, *Machine Learning*, *50*(1-2), 95–125, friedman, N Koller, D.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er (2000), Using bayesian networks to analyze expression data, *Journal of computational biology*, *7*(3-4), 601–620.

Friguet, C., M. Kloareg, and D. Causeur (2009), A factor model approach to multiple testing under dependence, *Journal of the American Statistical Association*, *104*(488), 1406–1415.

Fu, F., and Q. Zhou (2013), Learning sparse causal gaussian networks with experimental intervention: regularization and coordinate descent, *Journal of the American Statistical Association*, *108*(501), 288–300.

Gardner, T. S., D. di Bernardo, D. Lorenz, and J. J. Collins (2003), Inferring genetic networks and identifying compound mode of action via expression profiling, *Science*, *301*(5629), 102–105.

Gevins, A., and M. E. Smith (2000), Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style, *Cerebral Cortex*, *10*(9), 829–839.

Gibson, G. (2008), The environmental contribution to gene expression profiles, *Nature Reviews Genetics*, *9*(8), 575–581.

Granger, C. W. (1969), Investigating causal relations by econometric models and cross-spectral methods, *Econometrica: Journal of the Econometric Society*, pp. 424–438.

Grzebyk, M., P. Wild, and D. Chouanire (2004), On identification of multifactor models with correlated residuals, *Biometrika*, *91*(1), 141–151.

Hansen, L. P. (1982), Large sample properties of generalized-method of moments estimators, *Econometrica*, *50*(4), 1029–1054.

Harris, D., L. Mtys, D. Harris, and L. Mtys. (1999), *Introduction to the Generalized Method of Moments Estimation Generalized Method of Moments Estimation*, Cambridge University Press.

Hartemink, A. J., D. K. Gifford, T. S. Jaakkola, and R. A. Young (), Combining location and expression data for principled discovery of genetic regulatory network models, in *Pacific symposium on biocomputing*, vol. 7, pp. 437–449.

Heckerman, D., D. Geiger, and D. M. Chickering (1995), Learning bayesian networks - the combination of knowledge and statistical-data, *Machine Learning*, *20*(3), 197–243, heckerman, d geiger, d chickering, dm.

Higham, N. J. (1988), Computing a nearest symmetric positive semidefinite matrix, *Linear Algebra and Its Applications*, *103*, 103–118.

Hirose, K., and S. Konishi (2012), Variable selection via the weighted group lasso for factor analysis models, *Canadian Journal of Statistics*, *40*(2), 345–361.

Holst, F., et al. (2007), Estrogen receptor alpha (esr1) gene amplification is frequent in breast cancer, *Nature Genetics*, *39*(5), 655–660.

Horlings, H. M., et al. (2010), Integration of dna copy number alterations and prognostic gene expression signatures in breast cancer patients, *Clinical Cancer Research*, *16*(2), 651–663.

Hu, Y. N., and P. X. K. Song (2012), Sample size determination for quadratic inference functions in longitudinal design with dichotomous outcomes, *Statistics in Medicine*, *31*(8), 787–800.

Huang, J., N. Liu, M. Pourahmadi, and L. Liu (2006), Covariance matrix selection and estimation via penalised normal likelihood, *Biometrika*, *93*(1), 85–98.

Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai (2001), Lethality and centrality in protein networks, *Nature*, *411*(6833), 41–42.

Johnson, R. A., and D. W. Wichern (2007), *Applied multivariate statistical analysis*, 6th ed., xviii, 773 p. pp., Pearson Prentice Hall, Upper Saddle River, N.J.

Kaiser, H. F. (1958), The varimax criterion for analytic rotation in factor-analysis, *Psychometrika*, *23*(3), 187–200.

Kalisch, M., and P. Bhlmann (2013), Causal structure learning and inference: A selective review.

Kalisch, M., and P. Buhlmann (2007), Estimating high-dimensional directed acyclic graphs with the pc-algorithm, *Journal of Machine Learning Research*, *8*, 613–636, kalisch, Markus Buehlmann, Peter.

Kooperberg, C., M. LeBlanc, and V. Obenchain (2010), Risk prediction using genomewide association studies, *Genetic epidemiology*, *34*(7), 643–652.

Kustra, R., R. Shioda, and M. Zhu (2006), A factor analysis model for functional genomics, *Bmc Bioinformatics*, *7*.

Lahti, L., M. Schafer, H. U. Klein, S. Bicciato, and M. Dugas (2013), Cancer gene prioritization by integrative analysis of mrna expression and dna copy number data: a comparative review, *Briefings in Bioinformatics*, *14*(1), 27–35.

Lam, W., and F. Bacchus (1994), Learning bayesian belief networks: An approach based on the mdl principle, *Computational intelligence*, *10*(3), 269–293.

Ledoit, O., and M. Wolf (2004), A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis*, *88*(2), 365–411.

Lee, W., and Y. F. Liu (2012), Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood, *Journal of Multivariate Analysis*, *111*, 241–255.

Leek, J. T., and J. D. Storey (2007), Capturing heterogeneity in gene expression studies by surrogate variable analysis, *Plos Genetics*, *3*(9), 1724–1735.

Levina, E., A. Rothman, and J. Zhu (2008), Sparse estimation of large covariance matrices via a nested lasso penalty, *The Annals of Applied Statistics*, *2*(1), 245–263.

Li, F., and Y. Yang (2004), Recovering genetic regulatory networks from micro-array data and location analysis data, *GENOME INFORMATICS SERIES*, *15*(2), 131.

Li, F., and Y. Yang (2005), Using modified lasso regression to learn large undirected graphs in a probabilistic framework, in *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, vol. 20, p. 801, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Liang, K. Y., and S. L. Zeger (1986), Longitudinal data-analysis using generalized linear-models, *Biometrika*, *73*(1), 13–22.

Liu, H., F. Han, M. Yuan, J. Lafferty, and L. Wasserman (2012), High-dimensional semiparametric gaussian copula graphical models, *Annals of Statistics*, *40*(4), 2293–2326.

Lopes, C. T., M. Franz, F. Kazi, S. L. Donaldson, Q. Morris, and G. D. Bader (2010), Cytoscape web: an interactive web-based network browser, *Bioinformatics*, *26*(18), 2347–2348.

Lozoff, B., and M. K. Georgieff (2006), Iron deficiency and brain development, in *Seminars in pediatric neurology*, vol. 13, pp. 158–165, Elsevier.

Lucas, J. E., H.-N. Kung, and J.-T. A. Chi (2010), Latent factor analysis to discover pathway-associated putative segmental aneuploidies in human cancers, *PLoS computational biology*, *6*(9), e1000,920.

Lutz, R. W., and P. Buhlmann (2006), Boosting for high-multivariate responses in high-dimensional linear regression, *Statistica Sinica*, *16*(2), 471–494.

Mai, X. Q., L. Xu, M. Y. Li, J. Shao, Z. Y. Zhao, R. A. deRegnier, C. A. Nelson, and B. Lozoff (2012), Auditory recognition memory in 2-month-old infants as assessed by event-related potentials, *Developmental Neuropsychology*, *37*(5), 400–414.

McLean, E., M. Cogswell, I. Egli, D. Wojdyla, and B. de Benoist (2009), Worldwide prevalence of anaemia, who vitamin and mineral nutrition information system, 19932005, *Public health nutrition*, *12*(04), 444–454.

Meinshausen, N., and P. Buehlmann (2006), High-dimensional graphs and variable selection with the lasso, *Annals of Statistics*, *34*(3), 1436–1462.

Mootha, V. K., et al. (2003), Pgc-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nature Genetics*, *34*(3), 267–273.

Mosca, E., R. Alfieri, I. Merelli, F. Viti, A. Calabria, and L. Milanesi (2010), A multilevel data integration resource for breast cancer study, *Bmc Systems Biology*, *4*.

Newey, W. K. (1991), Uniform-convergence in probability and stochastic equicontinuity, *Econometrica*, *59*(4), 1161–1167.

Newman, M. (2010), *Networks: An Introduction*, 720 pp., Oxford University Press, Inc.

Onatski, A. (2009), Testing hypotheses about the number of factors in large factor models, *Econometrica*, *77*(5), 1447–1479.

Onatski, A. (2010), Determining the number of factors from empirical distribution of eigenvalues, *The Review of Economics and Statistics*, *92*(4), 1004–1016.

Pan, W. (2001), Akaike's information criterion in generalized estimating equations, *Biometrics*, *57*(1), 120–125.

Pearl, J. (2000), *Causality: models, reasoning and inference*, vol. 29, Cambridge Univ Press.

Peer, D., A. Regev, G. Elidan, and N. Friedman (2001), Inferring subnetworks from perturbed expression profiles, *Bioinformatics*, *17*(suppl 1), S215–S224.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009), Partial correlation estimation by joint sparse regression models, *Journal of the American Statistical Association*, *104*(486), 735–746.

Peng, J., J. Zhu, A. Bergamaschi, W. Han, D. Y. Noh, J. R. Pollack, and P. Wang (2010), Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer, *Annals of Applied Statistics*, *4*(1), 53–77.

Pollack, J. R., C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown (1999), Genome-wide analysis of dna copy-number changes using cdna microarrays, *Nature Genetics*, *23*(1), 41–46.

Pollack, J. R., et al. (2002), Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors, *Proceedings of the National Academy of Sciences of the United States of America*, *99*(20), 12,963–12,968.

Pourahmadi, M. (1999), Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation, *Biometrika*, *86*(3), 677–690, times Cited: 154.

Qu, A., and R. Z. Li (2006), Quadratic inference functions for varying-coefficient models with longitudinal data, *Biometrics*, *62*(2), 379–391.

Qu, A., and B. G. Lindsay (2003), Building adaptive estimating equations when inverse of covariance estimation is difficult, *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *65*, 127–142.

Qu, A., B. G. Lindsay, and B. Li (2000), Improving generalised estimating equations using quadratic inference functions, *Biometrika*, *87*(4), 823–836.

Qu, A., J. J. Lee, and B. G. Lindsay (2008), Model diagnostic tests for selecting informative correlation structure in correlated data, *Biometrika*, *95*(4), 891–905.

Qu, A., B. G. Lindsay, and L. Lu (2010), Highly efficient aggregate unbiased estimating functions approach for correlated data with missing at random, *Journal of the American Statistical Association*, *105*(489), 194–204.

Radlowski, E. C., and R. W. Johnson (2013), Perinatal iron deficiency and neurocognitive development, *Frontiers in human neuroscience*, *7*.

Ranade, S. C., S. Nawaz, A. Chakrabarti, P. Gressens, and S. Mani (2013), Spatial memory deficits in maternal iron deficiency paradigms are associated with altered glucocorticoid levels, *Hormones and behavior*, *64*(1), 26–36.

Richardson, T., and P. Spirtes (2002), Ancestral graph markov models, *Annals of Statistics*, pp. 962–1030.

Robinson, R. W. (1973), *Counting Labeled Acyclic Digraphs*, pp. 239–273, Academic Press, New York.

Rothman, A. J., E. Levina, and J. Zhu (2010), Sparse multivariate regression with covariance estimation, *Journal of Computational and Graphical Statistics*, *19*(4), 947–962.

Rubin, D. B., and D. T. Thayer (1982), Em algorithms for ml factor-analysis, *Psychometrika*, *47*(1), 69–76.

Sachs, K., O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan (2005), Causal protein-signaling networks derived from multiparameter single-cell data, *Science*, *308*(5721), 523–529.

Schfer, J., and K. Strimmer (2005), An empirical bayes approach to inferring large-scale gene association networks, *Bioinformatics*, *21*(6), 754–764.

Schmidt, M. (2010), Graphical model structure learning with l1-regularization, Ph.D. thesis.

Schneeweiss, H., and H. Mathes (1995), Factor-analysis and principal components, *Journal of Multivariate Analysis*, *55*(1), 105–124.

Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman (2003), Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nature genetics*, *34*(2), 166–176.

Sharan, S. K., et al. (1997), Embryonic lethality and radiation hypersensitivity mediated by rad51 in mice lacking brca2, *Nature*, *386*(6627), 804–810.

Shimamura, T., S. Imoto, R. Yamaguchi, and S. Miyano (2007), Weighted lasso in graphical gaussian modeling for large gene network estimation based on microarray data, *Genome Informatics*, *19*(142).

Shojaie, A., and G. Michailidis (2010), Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs, *Biometrika*, *97*(3), 519–538.

Siddappa, A. M., M. K. Georgieff, S. Wewerka, C. Worwa, C. A. Nelson, and R. A. DeRegnier (2004), Iron deficiency alters auditory recognition memory in newborn infants of diabetic mothers, *Pediatric Research*, *55*(6), 1034–1041.

Sieberts, S. K., and E. E. Schadt (2007), Moving toward a system genetics view of disease, *Mammalian Genome*, *18*(6-7), 389–401.

Snchez, B. N., E. Budtz-Jrgensen, L. M. Ryan, and H. Hu (2005), Structural equation models: a review with applications to environmental epidemiology, *Journal of the American Statistical Association*, *100*(472), 1443–1455.

Song, X. (2007), *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer.

Spirtes, P., C. N. Glymour, and R. Scheines (2000), *Causation, prediction, and search*, vol. 81, MIT press.

Stegle, O., A. Kannan, R. Durbin, and J. Winn (2008), *Accounting for non-genetic factors improves the power of eQTL studies*, *Lecture Notes in Bioinformatics*, vol. 4955, pp. 411–422.

Stein, C. (1956), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Third Berkeley Symposium on Mathematical Statistics and Probability, pp. 197–206, University of California Press.

Subramanian, A., et al. (2005), Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15,545–15,550.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005), Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *67*, 91–108.

Tipping, M. E., and C. M. Bishop (1999), Probabilistic principal component analysis, *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *61*, 611–622.

Tsamardinos, I., L. E. Brown, and C. F. Aliferis (2006), The max-min hill-climbing bayesian network structure learning algorithm, *Machine learning*, *65*(1), 31–78.

Tseng, S., P.; Yun (2009), A block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization, *Journal of Optimization Theory and Applications*, *140*(3), 513–535.

Turlach, B. A., W. N. Venables, and S. J. Wright (2005), Simultaneous variable selection, *Technometrics*, *47*(3), 349–363.

Van de Geer, S., and P. Bhlmann (2013), ell_0-penalized maximum likelihood for sparse directed acyclic graphs, *The Annals of Statistics*, *41*(2), 536–567.

Van Wieringen, W. N., and M. A. Van De Wiel (2011), Exploratory factor analysis of pathway copy number data with an application towards the integration with gene expression data, *Journal of Computational Biology*, *18*(5), 729–741.

Wall, M., A. Rechtsteiner, and L. Rocha (2003), Singular value decomposition and principal component analysis., *A practical approach to microarray data analysis*, pp. 91–109.

Wang, S. D., T. S. Kuo, and C. F. Hsu (1986), Trace bounds on the solution of the algebraic matrix riccati and lyapunov equation, *Ieee Transactions on Automatic Control*, *31*(7), 654–656.

Wang, Y. G., and V. Carey (2003), Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance, *Biometrika*, *90*(1), 29–41.

WHO (1999), Prevention and control of iron deficiency anaemia in women and children.

Wille, A., et al. (2004), Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana, *Genome Biol*, *5*(11), R92.

Wu, T. T., Y. F. Chen, T. Hastie, E. Sobel, and K. Lange (2009), Genome-wide association analysis by lasso penalized logistic regression, *Bioinformatics*, *25*(6), 714–721.

Wu, W. B., and M. Pourahmadi (2003), Nonparametric estimation of large covariance matrices of longitudinal data, *Biometrika*, *90*(4), 831–844.

Xiong, M., J. Li, and X. Fang (2004), Identification of genetic networks, *Genetics*, *166*(2), 1037–1052.

Yang, C., L. Wang, S. Q. Zhang, and H. Y. Zhao (2013), Accounting for non-genetic factors by low-rank representation and sparse regression for eqtl mapping, *Bioinformatics*, *29*(8), 1026–1034.

Yang, J., L. Li, and A. Wang (2011), A partial correlation-based bayesian network structure learning algorithm under linear sem, *Knowledge-Based Systems*, *24*(7), 963–976.

Yin, J. X., and H. Z. Li (2011), A sparse conditional gaussian graphical model for analysis of genetical genomics data, *Annals of Applied Statistics*, *5*(4), 2630–2650.

Yuan, M., and Y. Lin (2006), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *68*, 49–67.

Yuan, M., and Y. Lin (2007), Model selection and estimation in the gaussian graphical model, *Biometrika*, *94*(1), 19–35.

Yuan, Y. Y., C. Curtis, C. Caldas, and F. Markowetz (2012), A sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes, *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, *9*(4), 947–954.

Zhou, J., and A. Qu (2012), Informative estimation and selection of correlation structure for longitudinal data, *Journal of the American Statistical Association*, *107*(498), 701–710.

Zhou, Q. (2011), Multi-domain sampling with applications to structural inference of bayesian networks, *Journal of the American Statistical Association*, *106*(496), 1317–1330, zhou, Qing.

Zou, H. (2006), The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zou, H., and T. Hastie (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *67*, 301–320.

Zou, H., T. Hastie, and R. Tibshirani (2006), Sparse principal component analysis, *Journal of computational and graphical statistics*, *15*(2), 265–286.